

Doble grado Ingeniería Informática y Administración de Empresas.

(2019-2020)

Trabajo Fin de Grado

“Sistema de detección automática de emociones de
comunidades deportivas en textos de redes sociales”

Andrea Sánchez Muñoz

Tutor

José María Álvarez Rodríguez

Colmenarejo, julio de 2020



Esta obra se encuentra sujeta a la licencia Creative Commons **Reconocimiento – No Comercial – Sin Obra**

Agradecimientos

A mi tutor Chema por haber confiado en mí para realización de este interesante proyecto de la cátedra y por ayudarme, guiarme y motivarme durante el desarrollo del mismo.

A todo el equipo *The Reuse Company* el cual me ha tratado y ayudado de forma increíble, haciendo que sea un placer trabajar con ellos. En especial a mis compañeros de la cátedra Carlos, Cristina y Edu.

A mi familia quien ha sido mi principal apoyo en estos meses de confinamiento y durante toda la carrera.

A mis amigos y compañeros de carrera con los que he compartido momentos inolvidables y también han supuesto un apoyo fundamental durante mi etapa universitaria.

A la Universidad Carlos III y su personal docente y administrativo por su trabajo y dedicación.

Resumen Ejecutivo

El auge de las redes sociales hoy en día las ha situado como uno de los principales medios de comunicación donde los usuarios expresan sus pensamientos, sentimientos u opiniones e interactúan con otros usuarios y comunidades. Uno de los ámbitos que genera en redes sociales una alta actividad de carácter emocional es el deportivo, por lo que este medio supone una fuente de información de alto valor para conocer la opinión o el estado emocional de sus seguidores o aficionados. Bajo la oportunidad de explotar dicha información surge el objetivo del presente proyecto, el cual consiste en el desarrollo de un sistema de detección automática de la emoción expresada en redes sociales de ciertas comunidades o aficionados durante el transcurso de eventos deportivos a través del análisis de sus publicaciones.

Este proyecto surge dentro del acuerdo de colaboración entre RTVE y la UC3M, que conforma la cátedra RTVE-UC3M y que tiene como fin la investigación en el análisis de redes sociales y su aplicación en el sector audiovisual.

El sistema propuesto pretende detectar la emoción expresada por aquellos seguidores y comunidades de interés, cuyas cuentas son monitorizadas, para así ofrecer dicha información a RTVE como soporte para la toma de decisiones o personalización de contenidos. El sistema, pese a ser diseñado para ser aplicado a diferentes ámbitos, tendrá como primer alcance el análisis de la emoción asociada a eventos deportivos en la liga 2B grupo IV de fútbol ya que dichos eventos no se encuentran recogidos en la información deportiva proporcionada actualmente por RTVE.

En el presente documento se describe el análisis, diseño e implementación así como la evaluación del sistema propuesto, el cual emplea para el análisis de emociones un enfoque basado en conocimiento, conformado por una ontología y una serie de patrones, y ofrece al usuario una interfaz mediante la cual puede interactuar para la selección del equipo y jornada a analizar así como visualizar gráficamente la emoción global inferida y las publicaciones asociadas a dicha emoción.

Palabras clave: detección de emociones, redes sociales, procesamiento del lenguaje natural, léxico de emociones, base de conocimiento, *pattern matching*.

Índice

1.	Introducción.....	1
1.1.	La cátedra de RTVE-UC3M.....	2
1.1.1.	Proyecto Social Media News.....	3
1.2.	Motivación.....	5
1.3.	Objetivos.....	6
1.4.	Glosario.....	8
1.4.1.	Acrónimos	8
1.4.2.	Definiciones.....	9
1.5.	Estructura del documento	10
2.	Estado del Arte	12
2.1.	Representación de las emociones.....	13
2.2.	Procesamiento del lenguaje natural	15
2.2.1.	Técnicas de procesamiento del lenguaje natural	15
2.2.2.	Herramientas para el análisis de texto	17
2.2.3.	Problemáticas en la detección de emociones en texto.....	18
2.2.4.	Recursos para la detección de emociones.....	19
2.3.	Metodologías para la detección de emociones.....	20
2.4.	Tecnologías para la detección de emociones.....	24
3.	Análisis del sistema	25
3.1.	Análisis del problema.....	25
3.1.1.	Definición del problema de ingeniería	25
3.1.2.	Definición de la solución propuesta	25
3.2.	Especificación de requisitos.....	27

3.2.1.	Requisitos funcionales	28
3.2.2.	Requisitos no funcionales	35
3.3.	Diseño de casos de uso	39
3.3.1.	Diagrama de casos de uso	40
3.3.2.	Definición de casos de uso	41
4.	Diseño del Sistema	49
4.1.	Infraestructura tecnológica	49
4.1.1.	Microsoft Visual Studio	49
4.1.2.	C#	50
4.1.3.	Angular	50
4.1.4.	ArangoDB	51
4.1.5.	MySQL	51
4.1.6.	Knowledge Manager	52
4.2.	Definición de la Arquitectura del sistema	52
4.2.1.	Vista lógica	54
4.2.2.	Vista de Desarrollo	58
4.3.	Proceso de detección de emociones en los tweets	74
4.4.	Ontología que conforma la base de conocimiento	76
5.	Implementación	83
5.1.	Módulo de obtención de recursos	83
5.2.	Módulo de construcción de la ontología: configuración del componente Knowledge Manager	87
5.3.	Módulo de obtención de contexto objetivo	91
5.4.	Módulo análisis y clasificación de emociones	93
5.5.	Submódulo de Social Media News de muestra de resultados	97

6.	Pruebas	99
6.1.	Pruebas de validación	99
6.2.	Evaluación de los resultados de la clasificación	102
7.	Marco Legislativo	107
8.	Entorno Socioeconómico	108
9.	Planificación	110
9.1.	Fases del proyecto	110
9.2.	Diagrama de Gantt	111
10.	Presupuesto	112
10.1.	Coste de personal	112
10.2.	Coste de material	114
10.2.1.	Costes de Hardware	114
10.2.2.	Costes de Software	114
10.3.	Costes Indirectos	115
10.4.	Coste Total del Proyecto	115
11.	Conclusiones y Trabajo Futuro	116
11.1.	Conclusiones	116
11.2.	STrabajo Futuro	117
12.	Extended Abstract	119
12.1.	Abstract	119
12.2.	Introduction	120
12.3.	Motivation	121
12.4.	Objectives	121
12.5.	Emotion detection	123
12.5.1.	Methodologies for emotion detection	123

12.6.	Proposed system.....	125
12.6.1.	Emotion detection process in tweets.....	125
12.7.	Experiments and results	127
12.8.	Conclusions	128
	Bibliografia.....	130

Índice de Tablas

Tabla 1. Ejemplo tabla de requisitos.....	27
Tabla 2. RF-01 Extracción de contexto objetivo	28
Tabla 3. RF-02 Almacenamiento de contexto objetivo.....	29
Tabla 4. RF-03 Consulta de contexto objetivo	29
Tabla 5. RF-04 Monitorización de cuentas de usuarios en redes sociales.....	30
Tabla 6. RF-05 Filtrado de recursos	30
Tabla 7. RF-06 Eliminación de ruido en recursos	31
Tabla 8. RF-07 Estrategias de clasificación de emociones.....	31
Tabla 9. RF-08 Clasificación emocional de los recursos	32
Tabla 10. RF-09 Construcción de ontología.....	32
Tabla 11. RF-10 Determinación de la emoción global en función del contexto objetivo ...	32
Tabla 12. RF-11 Servicio web de consulta.....	33
Tabla 13. RF-12 Resultado servicio web.....	33
Tabla 14. RF-13 Almacenamiento de observaciones	33
Tabla 15. RF-14 Selección de equipo.....	34
Tabla 16. RF-15 Selección de jornada.....	34
Tabla 17. RF-16 Visualización de la emoción inferida	34
Tabla 18. RNF-01 Tiempo de respuesta servicio.....	35
Tabla 19. RNF-02 Tiempo de carga de resultados en Interfaz.....	35
Tabla 20. RNF-03 Integración del sistema.....	36
Tabla 21. RNF-04 Resultado de análisis en web.....	36
Tabla 22. RNF-05 Integración con servicio web proveedor de información deportiva.....	36
Tabla 23. RNF-06 Framework de desarrollo.....	37
Tabla 24. RNF-07 Tipo de servicio web.....	37
Tabla 25. RNF-08 Base de datos ontología.....	37
Tabla 26. RNF-09 Base de datos observaciones.....	38
Tabla 27. RNF-10 Fuente de información deportiva.....	38
Tabla 28. RNF-11 Formato datos información deportiva.....	38
Tabla 29. RNF-12 Base de datos información deportiva.....	39

Tabla 30. RNF-13 Periodicidad de generación de observaciones.	39
Tabla 31. Tabla ejemplo de descripción de casos de uso	41
Tabla 32. CU-01 Selección parámetros	43
Tabla 33. CU-02 Detección emociones	44
Tabla 34. CU-03 Visualización emoción inferida	45
Tabla 35. CU-04 Consulta información deportiva	46
Tabla 36. CU-05 Monitorización Twitter	47
Tabla 37. CU-06 Configuración ontología	48
Tabla 38. Componente Interfaz	61
Tabla 39. Componente NewsController	63
Tabla 40. Componente EmotionController.....	66
Tabla 41. Componente NewsExtractor.....	69
Tabla 42. Componente EmotionExtractor	73
Tabla 43. Componente Cake.....	74
Tabla 44. Distribución de términos por emoción	78
Tabla 45. Patrones para la detección de negación.	80
Tabla 46. Patrones para la detección de intensificadores.	81
Tabla 47. Patrones para la detección de mitigadores.....	82
Tabla 48. Exponentes según resultado de función exponencial para emoción positiva.	96
Tabla 49. Exponentes según resultado de función exponencial para emoción negativa.	96
Tabla 50. Tabla ejemplo de definición de prueba de validación.	99
Tabla 51. PR-01 Obtención y almacenamiento contexto objetivo	100
Tabla 52. PR-02 Consulta contexto objetivo.....	100
Tabla 53. PR-03 Monitorización de cuentas y hashtags.....	100
Tabla 54. PR-04 Consulta de recursos.....	101
Tabla 55. PR-05 Detección de emoción de los recursos.	101
Tabla 56. PR-06. Consulta ontología.....	101
Tabla 57. PR-07 Consulta del resultado del análisis.	102
Tabla 58. PR-08 Visualización del resultado del análisis.	102
Tabla 59. Cálculo métricas primera evaluación de clasificación	104
Tabla 60. Cálculo métricas segunda evaluación de clasificación.....	105

Tabla 61. Ejemplos resultado de la clasificación de emociones.....	106
Tabla 62. Planificación de fases y asignación de recursos del proyecto.	110
Tabla 63. Cálculo coste personal estimado por hora trabajada	113
Tabla 64. Cálculo del coste total estimado de personal.....	113
Tabla 65. Costes estimados de material hardware.....	114
Tabla 66. Costes estimados de material software.....	114
Tabla 67. Presupuesto estimado del proyecto.	115

Índice de Figuras

Figura 1. Arquitectura de alto nivel del sistema Social Media News	4
Figura 2. Rueda de emociones de Plutchik [11].....	14
Figura 3. Modelo circunflejo de las Emociones de J.A. Russell [12]	15
Figura 4. Diagrama de casos de uso	40
Figura 5. Arquitectura de alto nivel del sistema Social Media News (Módulo análisis)	53
Figura 6. Diagrama de clases.....	54
Figura 7. Diagrama de componentes	59
Figura 8. Esquema general de la arquitectura funcional del sistema de análisis.....	75
Figura 9. Ejemplo de Spanish Emotional Lexicon (SEL) [26]	77
Figura 10. Clasificación de emojis por emoción [22].	78
Figura 11. Cuentas y Hashtags de Twitter monitorizados.....	84
Figura 12. Pantalla de monitorización de cuentas de la plataforma SMART-RTVE.	85
Figura 13. Colecciones para el almacenamiento de recursos e información de usuario de ArangoDB.....	85
Figura 14. Ejemplo de recurso en formato JSON.....	86
Figura 15. Interfaz de Knowledge Manager para insertar lista de términos desde Excel. ..	88
Figura 16. Interfaz de Knowledge Manager que muestra clúster <<Alegría>>	88
Figura 17. Interfaz de Knowledge Manager que muestra clúster <<Jerga Alegría>>	89
Figura 18. Interfaz de Knowledge Manager que muestra clúster <<Eventos no deportivos>>	90
Figura 19. Interfaz de Knowledge Manager que muestra clúster <<Otras categorías fútbol >>	90
Figura 20. Interfaz de Knowledge Manager que muestra la configuración de patrones.	91
Figura 21. Colección para el almacenamiento de información deportiva en ArangoDB....	92
Figura 22. Ejemplo de observación de emoción global en formato JSON	95
Figura 23. Ejemplo de observación de porcentajes de emociones en formato JSON	95
Figura 24. Componente Interfaz para la selección de parámetros.....	98
Figura 25. Componente Interfaz para la visualización de resultados.....	98
Figura 26. Diagrama de Gantt	111

Figura 27. General scheme of the functional architecture of the analysis system. 126

1. Introducción

Las redes sociales hoy en día han dejado de ser una moda y se han convertido en uno de los principales medios de comunicación en masa, situándose en una etapa de ciclo de vida de maduración. Mediante las redes sociales el ser humano es capaz de satisfacer una de sus principales necesidades que es la de comunicarse, expresando sus pensamientos, sentimientos u opiniones e interactuando con otros usuarios y comunidades. Por ello, en este medio se generan una gran cantidad de publicaciones que suponen una fuente de información de alto valor para su estudio en multitud de áreas de conocimiento, como puede ser la psicología, sociología, política, marketing, informática, etc.

Desde un punto de vista analítico, uno de los campos que obtienen un gran provecho de esa cantidad ingente de información, en forma de publicaciones y/o comentarios, es el análisis de sentimiento, opiniones o emociones. El uso de datos públicamente disponibles para llevar a cabo un análisis emocional reduce de forma significativa los costes, esfuerzos y el tiempo necesario para la gestión de encuestas públicas de gran escala y cuestionarios utilizados para la obtención de la información necesaria para dicho análisis.

Para que sea posible el uso del gran volumen de información generado por las redes sociales, se requieren herramientas para la gestión de grandes cantidades de datos y su procesamiento de forma automática. En esta necesidad juegan un gran papel la tecnologías Big Data cuya proliferación resulta imparable en una era donde la transformación digital es la protagonista.

Las aplicaciones del análisis de sentimiento y emociones en redes sociales son múltiples. Por ejemplo, en el mundo del marketing se utiliza la opinión obtenida de los usuarios respecto a un producto o marca para el diseño de campañas personalizadas de marketing, la mejora en la comunicación con el cliente o la definición de estrategias de fidelización. Por otro lado, en el campo de la política se puede utilizar para extraer la opinión pública pudiendo así conocer el impacto de acciones políticas o realizar predicciones sobre futuras elecciones. Además de estos, existen otro gran número de posibles ejemplos.

Bajo este contexto surge el presente proyecto de la cátedra RTVE-UC3M, que dado el proceso de transformación digital llevado a cabo en el sector audiovisual y de medios de

comunicación pretende cambiar y mejorar la concepción de sus servicios actuales como los contenidos y la forma en que estos se proveen. Además, dicho proceso también busca una adaptación mayor a sus usuarios actuales y potenciales ofreciendo, por ejemplo, una mayor personalización de sus contenidos y servicios para los diferentes usuarios, grupos y comunidades.

Dada la plataforma SMART-RTVE, desarrollada por la cátedra RTVE-UC3M, que permite la monitorización de cuentas y *hashtags* en Twitter para la obtención de sus publicaciones, el presente proyecto busca cubrir la necesidad de análisis por parte de RTVE de recursos publicados por ciertos usuarios y/o comunidades para conocer sus emociones u opiniones respecto a diferentes temas, eventos, etc. Además, el sistema propuesto mostrará el resultado del análisis de una forma visual dentro de la plataforma mencionada. Las emociones u opiniones extraídas del análisis serán utilizadas por RTVE para la toma de decisiones, como por ejemplo la personalización del contenido de noticias en función de dichos parámetros.

En concreto, este proyecto buscará la obtención de las emociones presentes en la afición futbolística asociadas a partidos disputados. Por tanto, el sistema se centrará en la detección de emociones de aquellos recursos publicados durante la previa y el transcurso del partido por parte de cuentas y *hashtags* oficiales, de seguidores y de medios relacionados con los equipos de interés.

Sin embargo, con el sistema propuesto no se pretende resolver dicha necesidad de análisis únicamente para los eventos deportivos mencionados, sino que la propuesta busca un diseño general que pueda ser aplicable a otro tipo eventos deportivos.

1.1. La cátedra de RTVE-UC3M

La cátedra de RTVE-UC3M [1] se constituyó en el año 2015 como un acuerdo de colaboración entra la Fundación Carlos III y Radio Televisión Española para la realización de diferentes actividades de investigación, formación, divulgación y estudio dentro del campo de las Tecnologías de la Información y de las Comunicaciones aplicadas al mundo audiovisual. En concreto, dichas actividades se centran en las siguientes áreas de estudio:

- Linked Data y Big Data para la gestión de altos volúmenes de datos e información.

- Técnicas de análisis de información textual (por ejemplo, procesamiento del lenguaje natural) para la gestión de contenido.
- Análisis de sentimientos y emociones aplicado a contenido relacionado con el sector.

Actualmente, el principal foco de aplicación de las tecnologías mencionadas reside en las redes sociales. La cátedra surge de la necesidad por parte de los medios de comunicación de gestión de medios digitales y, en consecuencia, la investigación para la extracción de información de redes sociales supone uno de los principales objetivos.

En referencia al trabajo realizado por la cátedra y su foco hacia las redes sociales, cabe destacar la plataforma SMART-RTVE, cuyas siglas dentro del presente contexto significan Sistema de Monitorización y Alertas en la Red para Radio Televisión Española. Se trata de una plataforma distribuida y escalable que fue desarrollada para la obtención y almacenamiento de información procedente de redes sociales y su posterior monitorización mediante un sistema web.

Dada la capacidad ofrecida por dicha plataforma de obtención y monitorización de una gran cantidad de recursos, surge en la cátedra la necesidad de análisis de dichos datos y la aplicación de la información en el ámbito de la comunicación. De esta forma surge el proyecto Social Media News, que será explicado a continuación y del que forma parte el sistema de análisis de emociones propuesto en el presente documento.

1.1.1. Proyecto Social Media News

El proyecto Social Media News busca el enriquecimiento de noticias deportivas mediante la generación automática de texto que refleje el estado emocional presente en la afición durante el evento deportivo al que hace referencia la noticia, emoción que es extraída también de forma automática mediante el seguimiento y análisis de las redes sociales.

Los eventos deportivos objeto de este proyecto son la ligas inferiores de fútbol masculino. Se escoge este tipo de evento deportivo debido a que actualmente RTVE, en sus área de noticias deportivas, no recoge este tipo de información, por lo que la posibilidad de generación automática de crónicas deportivas con carácter emocional, tal como lo haría un periodista, resulta de gran interés para atraer al público seguidor de este tipo de evento y

aumentar en consecuencia el número de visitas en la web de noticias deportivas de RTVE. En un principio, el contenido deportivo del proyecto se centrará en la liga de segunda B grupo IV de fútbol

El sistema de análisis propuesto en el presente documento, como ya se ha comentado, formará parte del proyecto Social Media News, siendo el encargado de obtener la emoción presente entre los aficionados de los equipos estudiados mediante el análisis de los recursos que estos publican en la red social Twitter.

La arquitectura propuesta para el proyecto Social Media News, donde se definen los principales bloques funcionales, es la siguiente:



Figura 1. Arquitectura de alto nivel del sistema Social Media News

- **Adquisición de datos:** módulo encargado de la conexión con la red social Twitter y su escucha activa. Esta funcionalidad será llevada a cabo por la plataforma ya existente SMART-RTVE.
- **Escalabilidad:** controlador encargado de dirigir los diferentes procesos del sistema.
- **Almacenamiento:** módulo encargado del almacenamiento y consulta de la información del sistema.

- **Análisis:** módulo que recoge las diferentes tareas de análisis de los datos sistema . Los submódulos que contienen son el gestor de ontologías que se encarga de la conexión con la herramienta de gestión del conocimiento utilizada para el análisis, el gestor del lenguaje natural que realiza las tareas de procesamiento del lenguaje natural aplicadas a los recursos del sistema, el gestor de confianza que se encarga de asegurar la obtención de información relevante para el sistema y, por último, el gestor de emociones que lleva a cabo la clasificación de los recursos en función de la emoción que expresan.
- **Servicio de generación de crónicas deportivas:** bloque encargado de la generación de texto correspondiente a crónicas deportivas enriquecidas con la emoción inferida del módulo de análisis.
- **Aplicación cliente:** capa de presentación de la aplicación para la visualización de los resultados obtenidos del sistema, tanto de la noticia generada como de las emociones extraídas del módulo de análisis. Además, permite la interacción del usuario para la configuración de los parámetros del sistema.

1.2. Motivación

Los eventos deportivos es un tema que normalmente genera debate y en la información deportiva publicada en los medios se puede percibir el carácter crítico derivado de dicha controversia. Hoy en día, existen multitud de portales de datos abiertos, como puede ser LFP o NBA.com, desde los cuales se obtienen multitud de información deportiva, incluso en algunos casos con cierto carácter crítico. Sin embargo, las redes sociales, mediante las publicaciones de comentarios o crónicas sobre eventos deportivos, son uno de los medios donde mejor se observa dicho debate y las emociones expresadas por las diferentes comunidades ante un cierto resultado.

En los partidos de fútbol las redes sociales cuentan con un papel importante [2], en especial Twitter, produciéndose en esta red social un “partido paralelo”. Ante un partido de importancia (Champions, el clásico, etc.) desde la previa los aficionados comentan y debaten, mostrando así sus emociones y opiniones, llegando a hacer de este evento una tendencia en

Twitter. Por ejemplo, la final de Champions de 2016, disputada por el Real Madrid y el Atlético de Madrid, fue vista por televisión por casi 10 millones de personas y tuvo una audiencia a través de Twitter de 1.307.632 usuarios (aproximadamente un 15% respecto a la televisiva). El número de *tweets* publicados en España fueron de 1.200.000 *tweets*.

El fútbol cuenta con una gran afición en estos medios, como prueba de ello destaca la gran cantidad de seguidores de las cuentas oficiales de los equipos de fútbol y sus jugadores, estando en el *top 50* de Twitter en España gran cantidad de cuentas relacionadas con el deporte rey. Las cuatro primeras de dicho ranking son Cristiano con 78 millones, Neymarjr con 43 millones, Real Madrid con 32 millones y FCBarcelona con 30 millones. Además, otra prueba de la popularidad del fútbol en Twitter es el hecho de que lo *hashtags* de deporte más utilizados en esta red social son de carácter futbolístico, por ejemplo en 2017 el más utilizado fue #HalaMadrid.

Tal es el papel de la red social pájaro en el mundo futbolístico que en 2018, mediante un acuerdo con la FIFA, Twitter se convirtió en la plataforma homologada para que los emisores oficiales puedan reproducir los goles del campeonato.

Por tanto, es más que evidente que las redes sociales proporcionan una gran cantidad de información deportiva actualizada y con alta carga emocional que mediante un análisis de sentimiento, opinión o emoción de dichos datos se obtiene información de gran valor, resultando altamente representativa para el estudio estratégico del comportamiento social y la toma de acciones en consecuencia.

1.3. Objetivos

En base a las necesidades planteadas anteriormente, el objetivo global del presente proyecto es el de diseñar e implementar un sistema de detección automática de las emociones presentes entre la afición durante el transcurso de eventos deportivos mediante el análisis de sus publicaciones en redes sociales. Todo ello mediante el uso de técnicas de procesamiento de lenguaje natural y encaje de patrones cuyo resultado (la emoción inferida) será proporcionado al usuario de forma visual para su posterior utilización por el personal de RTVE para la toma de decisiones y acciones en consecuencia.

Para la consecución de dicho objetivo, se deben alcanzar los objetivos específicos enumerados a continuación:

1. Estudio de las metodologías y recursos para la detección automática de emociones y de los modelos emocionales existentes para la definición de las categorías de emociones en las que se va a basar la clasificación de recursos.
2. Análisis del sistema propuesto, especificando las diferentes funcionalidades que debe realizar, la interacción de los usuarios con este y las restricciones que debe cumplir.
3. Definición de una metodología para el tratamiento e integración de los recursos relativos a la información de interés provenientes de Twitter
4. Determinación del conjunto de dimensiones a identificar en la información deportiva necesario para llevar a cabo el análisis de emociones: evento deportivo, liga, grupo, equipo, fecha del evento, resultado, etc.
5. Diseño de una arquitectura adecuada del sistema que permita su futura integración y reutilización.
6. Configuración de una ontología y patrones semánticos adecuados para una correcta identificación de la emociones y de reglas de decisión /algoritmos para determinar la emoción (o la intensidad de esta) en función del contexto tanto subjetivo como objetivo.
7. Integración de la plataforma de monitorización de redes sociales SMART-RTVE y diseño e implementación del servicio de consulta de los recursos a analizar.
8. Implementación del servicio de análisis y despliegue de los algoritmos de clasificación de los recursos.
9. Implementación de una interfaz gráfica para la interacción y visualización de los resultados por parte del usuario.
10. Experimentación para la verificación y validación del correcto funcionamiento del sistema y evaluación de los resultados de la clasificación.

1.4. Glosario

1.4.1. Acrónimos

- **API:** *Application Programming Interface* (Interfaz de programación de aplicaciones)
- **AQL:** *ArangoDB Query Language* (Lenguaje de consulta ArangoDB)
- **IDE:** *Integrated Development Environment* (Entorno de desarrollo integrado)
- **JSON:** *JavaScript Object Notation* (Notación de objeto de JavaScript)
- **KM:** Knowledge Manager
- **LSTM:** *Long short-term memory* (memoria larga-corto plazo)
- **MVC:** Modelo Vista Controlador
- **NLP:** *Natural Language Processing* (Procesamiento del lenguaje natural)
- **OTRI:** Oficina de Transferencia de Resultados de Investigación.
- **POS:** *Part of speech* (Parte del discurso)
- **REST:** *Representational State Transfer* (Transferencia de Estado Representacional)
- **SEL:** *Spanish Emotional Lexicon* (Léxico emocional en español)
- **SMART-RTVE:** Sistema de Monitorización y Alertas en la Red para Radio Televisión Española
- **SQL:** *Structured Query Language* (Lenguaje de consulta estructurada)
- **SVM:** *Support Vector Machine* (Máquinas de vectores de soporte)
- **UML:** *Unified Modeling Language* (Lenguaje unificado de modelado)

1.4.2. Definiciones

- **Big Data:** conjunto de técnicas para el tratamiento y almacenamiento de gran cantidad de datos para los que los métodos de procesamiento convencionales no tienen capacidad de gestionarlos.
- **Clúster:** agrupación de los términos de una ontología para un cierto fin.
- **Dataset:** colección de datos estructurados.
- **Emojis:** imagen o pictograma utilizado en medios digitales para la expresión de un sentimiento, emoción o idea.
- **Framework:** marco de trabajo para la organización y desarrollo software.
- **Hashtag:** palabra o conjunto de palabras que aparecen precedidos del símbolo almohadilla (#) empleadas en los medios digitales para agrupar o etiquetar las publicaciones de acuerdo a una temática.
- **Keywords:** palabras clave para la búsqueda de información.
- **Ontología:** conjunto de datos que define la relación existente entre términos de un dominio que representan una base de conocimiento.
- **Pattern matching:** técnica para la detección de la presencia de los elementos pertenecientes a un patrón.
- **Slot:** ranura o hueco que representa los elementos que conforman un patrón.
- **Servicio web:** sistema software que permite la interacción interoperable a través de red de máquina a máquina.
- **Twitter:** red social basada en microblogging (variante de los blogs con textos de mayor brevedad) que permite la publicación y compartición de contenido en tiempo real.
- **Tweet:** unidad de texto con longitud máxima de 140 caracteres que corresponde a una publicación en la red social Twitter.

1.5. Estructura del documento

El presente documento se estructura en 11 capítulos que son explicados a continuación:

1. **Introducción:** breve explicación de la problemática planteada en el proyecto junto con la motivación de este y los objetivos a alcanzar.
2. **Estado del arte:** estudio de los principales trabajos y tecnologías existentes dentro del contexto del problema para la orientación de la solución a desarrollar.
3. **Análisis del sistema:** análisis del problema y justificación de la solución propuesta, especificación de los requisitos tanto funcionales como de restricción del sistema y descripción de los casos de usos que explican las acciones de los usuarios con el sistema.
4. **Diseño del sistema:** descripción detallada del diseño del sistema propuesto, explicando la infraestructura tecnológica empleada y especificando la arquitectura de este, tanto a alto nivel como siguiendo el modelo de vistas 4+1.
5. **Implementación:** descripción del proceso de implementación de los diferentes módulos que conforman el sistema propuesto.
6. **Pruebas:** especificación de las diferentes pruebas realizadas en el sistema y descripción de los resultados obtenidos.
7. **Marco legislativo:** estudio de los aspectos legales y normativos que pueden afectar al proyecto.
8. **Entorno Socioeconómico:** análisis del entorno de carácter socioeconómico en el que se encuadra el proyecto.
9. **Planificación:** descripción de las diferentes etapas del proyecto y las actividades que las conforman y especificación de la duración de estas para la planificación de desarrollo del proyecto.
10. **Presupuesto:** estimación de los costes asociados a la realización del proyecto.

11. Conclusiones y trabajos futuros: especificación de las conclusiones extraídas tras la realización del proyecto y el establecimiento de líneas futuras de trabajo.

2. Estado del Arte

En base a la problemática presentada, en este apartado se realiza el análisis del estado del arte relativo al proyecto para así estudiar las diferentes alternativas de soluciones existentes y, en consecuencia, establecer las bases tecnológicas sobre las que se va a sustentar el desarrollo del proyecto. Dado el objetivo principal del sistema, se tratarán las diferentes tecnologías y avances existentes en el campo de la detección de emociones en texto, en concreto en corpus de redes sociales.

En primer lugar, es necesario realizar una breve introducción al análisis de emociones en texto [3]. En el campo de la lingüística computacional, el análisis de emociones consiste en el proceso de clasificación de un texto en función de la emoción que expresa. Se puede entender como un estudio producto de la evolución natural del análisis de sentimientos, el cual consiste en la clasificación de un texto en cuanto a la polaridad de la carga subjetiva que presenta (positiva, negativa o neutral).

Cabe destacar, que el número de estudios y metodologías presentes en el campo del análisis de sentimientos es muy superior al de detección de emociones, contando con modelos refinados y resultados altamente satisfactorios. Este se trata de un campo ya establecido en el procesamiento del lenguaje natural en el que destaca su éxito en aplicaciones en el ámbito de la publicidad [4], resúmenes [5] o para la mejora del proceso de extracción de información en textos [6], entre otros.

A pesar de que existe un largo camino a recorrer para que el estudio de emociones alcance los resultados del análisis de sentimiento, la información adicional que proporciona puede suponer una alta mejora en las aplicaciones de este último. Por ejemplo, existen emociones que pese a tener una misma polaridad, como puede ser miedo e ira (polaridad negativa), estas pueden influir de diferente forma en el estado o comportamiento de los usuarios por lo que su distinción resulta de gran valor para su aplicación en campos como el marketing o la sociopolítica, entre otros.

Entre las diferentes aplicaciones en las que se ha empleado el análisis de emociones en los últimos años destacar que en marketing [7] se han analizado las reacciones de los

consumidores ante productos o servicios para tomar decisiones sobre qué aspectos se deben mejorar o cambiar respecto a estos para una mayor satisfacción del consumidor. Por otro lado, en línea con el cometido del presente proyecto, en el análisis de emociones en texto procedente de redes sociales se encuentran estudios, por ejemplo, para detectar desórdenes mentales como la depresión [8] a través del contenido que publican los usuarios en redes sociales o sobre el efecto de ciertos eventos de importancia [9] (de carácter social, económico, político, etc.) en el estado de ánimo de los usuarios, entre otras aplicaciones.

2.1. Representación de las emociones

Para la detección de emociones es necesario el estudio previo de los diferentes modelos emocionales existentes en el campo de la psicología, para así definir el modelo de referencia para la clasificación realizada por el sistema.

En primer lugar, cabe destacar que no existe un modelo emocional estándar, por lo que se estudiarán las teorías generales y de mayor aceptación en el campo de la psicología. En los estudios de este campo se distinguen principalmente 2 tipos de modelos: modelo de categorización discreta y modelo dimensional.

En primer lugar, la categorización discreta se basa en las teorías evolutivas que afirman que las emociones son biológicamente determinadas, discretas y pertenecen a unos pocos grupos, denominadas emociones básicas o fundamentales. El principal ejemplo es el modelo de las 6 emociones básicas de Ekman [10] donde se establece las siguientes categorías de emoción: tristeza, alegría, ira, miedo, aversión y sorpresa. Otro modelo que se usa como referencia es la rueda de Plutchik [11] el cual distingue 8 categorías básicas de emociones: alegría, confianza, miedo, sorpresa, tristeza, aversión, ira y anticipación. En este modelo dichas emociones se organizan en pares opuestos y se descomponen en emociones secundarias.

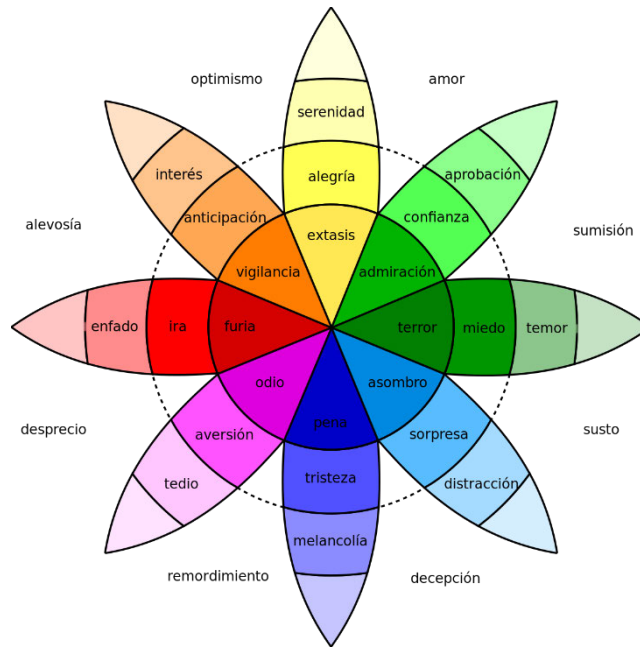


Figura 2. Rueda de emociones de Plutchik [11]

Por otro lado, los modelos dimensionales no establecen un número finito de emociones, sino que basan la clasificación en dimensiones. Esta perspectiva proviene de la hipótesis de que las emociones son resultado de un sistema neurofisiológico común e interconectado. Por tanto, los modelos basados en este punto de vista buscan un conjunto finito de características bajo las cuales se pueden descomponer las emociones, cuya combinación determina el estado afectivo. Un ejemplo a destacar es el modelo circuplejo de Russell [12] que distribuye las emociones en 2 dimensiones: placer o valencia, que indica como de positivo o negativo se siente una persona, y activación o estimulación, que mide como es de probable la toma de una acción en el estado de ánimo en el que se encuentra la persona.

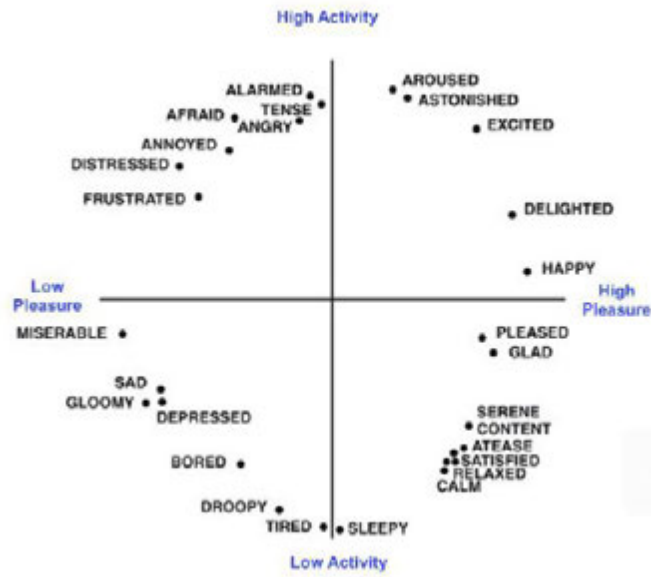


Figura 3. Modelo circunflejo de las Emociones de J.A. Russell [12]

En el campo del análisis de emociones, el tipo de modelo mayormente utilizado es el discreto por lo que resulta de preferencia para el presente proyecto. Además, ofrece una mayor usabilidad para la posterior aplicación en la toma de acciones o decisiones, como por ejemplo su uso para el enriquecimiento de noticias deportivas.

2.2. Procesamiento del lenguaje natural

En el campo de la detección de emociones en texto las técnicas de procesamiento del lenguaje natural tiene un papel fundamental. Esto se debe a que, antes de aplicar los algoritmos de clasificación del texto, es necesario un preprocesamiento de este para su transformación a un formato que permita el posterior tratamiento y clasificación de los datos de modo más sencillo y efectivo [13].

2.2.1. Técnicas de procesamiento del lenguaje natural

En este apartado se presentan las principales tareas de procesamiento del lenguaje empleadas en el análisis de texto:

- **Tokenización:** la técnica de tokenización consiste en la división del texto en *tokens*, que se tratan de instancias o unidades que componen el texto. Los principales niveles de división suelen ser frases y palabras. Esta división se realiza para poder llevar a

cabo un análisis individual de cada *token* y de la relación entre cada uno de los mismos. Cabe destacar que, en general, antes de emplear esta técnica es necesario llevar a cabo previamente ciertas tareas de limpieza de datos como son la eliminación de puntuaciones, caracteres repetidos, mayúsculas, etc.

- **Part of Speech (POS):** la técnica de *part of speech* consiste en la clasificación de cada *token* según la parte del discurso que le corresponda, es decir, su categoría morfosintáctica. Por tanto, como resultado de este proceso se obtiene cada *token* etiquetado con su categoría gramatical y otras características relacionadas con esta, como es el número, género o persona.
- **Lematización:** la técnica de lematización trata de cambiar cada palabra a su forma canónica o *lema*. Dicha transformación permite el tratamiento conjunto de todas las palabras flexionadas de un *lema*, pudiendo reducir el tamaño de la ontología o base de datos utilizada y mejorando la precisión en la recuperación de la información. Para llevar a cabo dicho proceso es necesario que se especifique la parte del discurso (POS) de la palabra ya que la transformación dependerá de si se trata de un adjetivo, nombre, verbo o adverbio. Por ejemplo, si se trata de un verbo, se transformarán las flexiones de las diferentes formas verbales al verbo en forma infinitiva. Por otro lado, cabe mencionar la técnica de *stemming*, que tiene similitudes con el proceso de lematización, pero en este caso la transformación de la palabra es hacia su raíz. Este proceso es mucho más rápido que la lematización, reduce en un grado muy superior el número de palabras en base de datos y permite encontrar relaciones entre palabras de distinto tipo. Sin embargo, el uso de *stemming* en el procesamiento de texto supone cierta pérdida de información semántica y sintáctica, aspecto que resulta decisivo en el campo de la detección de emociones, donde la información semántica del texto es de gran importancia, lo que conlleva a una preferencia por la lematización.
- **Desambiguación:** el proceso de desambiguación de palabras, en inglés *Word Sense Disambiguation* (WSD), consiste en un conjunto de técnicas para el tratamiento de conflictos de ambigüedad del sentido de una palabra debido a la polisemia del lenguaje. El objetivo de este proceso es la asignación automática del sentido correcto de las palabras tratadas. Las técnicas empleadas en este proceso pueden ser de distinto

tipo como el uso de restricciones morfosintácticas, métodos basados en conocimiento (recursos externos, diccionarios o tesauros) o el uso de técnicas de aprendizaje automático o estadísticas basadas en corpus de entrenamiento.

2.2.2. Herramientas para el análisis de texto

En este apartado se presentan una serie de herramientas para el procesamiento del lenguaje natural que permiten llevar a cabo las técnicas mencionadas en el subapartado anterior y requeridas para el posterior análisis de texto.

En primer lugar, una de las herramienta utilizadas para el análisis de texto en el ámbito de la detección de emociones es FreeLing [14]. Este software se trata de un conjunto de herramientas de código abierto creado por el grupo de investigación de la Universidad Politécnica de Barcelona sobre el procesamiento del lenguaje natural, proyecto dirigido por Lluís Padró.

FreeLing es una librería de C++ que ofrece diferentes funcionalidades para el procesamiento y análisis de texto en lenguaje natural, entre las cuales se pueden destacar el análisis morfológico, reconocimiento de entidades, etiquetado de parte del discurso (POS), desambiguación y etiquetado de los elementos de una oración según su rol semántico. Esta herramienta se ofrece en varios idiomas entre los que se encuentra el español y además puede ser utilizada a través de una línea de comandos *front-end* donde se puede escoger el formato de salida de los resultados (JSON, XML, etc.)

Por otro lado, otro software de uso común en este campo es NLKT [15] , plataforma de código abierto utilizada en programas en lenguaje Python que requieren del procesamiento de datos en lenguaje natural. NLKT proporciona un gran número de corpus de entrenamiento y recursos léxicos, como por ejemplo WordNet [16]. Además, también ofrece distintas librerías para el procesamiento de texto con las siguientes funcionalidades: tokenización, etiquetado, *stemming*, parseado y razonamiento semántico.

Por último, destacar la herramienta desarrollada por *The Reuse Company*, Knowledge Manager (KM) [17]. Esta se trata de un *software* que se encarga de la creación y gestión de bases de datos que se encuentran conformadas por un dominio de conocimiento (ontología)

y un conjunto de reglas lingüísticas propias de un idioma. La ontología incluye tanto vocabulario de términos del dominio como patrones que requieren de dichos términos. Entre las técnicas de procesamiento de lenguaje natural que son empleadas en el análisis de texto para su posterior clasificación emocional, KM ofrece las siguientes funcionalidades: lematización, tokenización, etiquetado POS, desambiguación y reconocimiento de patrones.

2.2.3. Problemáticas en la detección de emociones en texto.

La expresión de emociones en el lenguaje natural presenta una alta complejidad, existiendo ciertas problemáticas en el procesamiento del lenguaje natural que aumentan en alto grado la dificultad del análisis [3]. Cabe destacar problemáticas como las siguientes:

- **Negación:** la utilización de un conjunto de palabras con carga emotiva para la identificación de la emoción presenta la problemática de que si en un fragmento de texto aparece dicha palabra pero de forma negada y otra sin negar, es probable que se califique de la misma forma siendo su significado distinto o incluso contrario.
- **Metáforas, ironías, etc.:** el uso de figuras literarias en el lenguaje informal es altamente frecuente, lo que su comprensión para las computadoras resulta de gran complejidad.
- **Aparición de múltiples emociones:** un texto complejo puede presentar diversas emociones de diferente carácter por lo que determinar la emoción del conjunto resulta complicado.

Además, la expresión de las emociones en el lenguaje cuenta con una gran sensibilidad al contexto. Gran parte de los trabajos se basan en la expresión explícita de las emociones mediante la búsqueda de palabras con carga afectiva. Sin embargo, esto supone una limitación para la detección de las emociones ya que en un gran número de ocasiones estas son expresadas de forma implícita. El número de trabajos en la literatura de este campo sobre la identificación de la expresión implícita de emociones es reducido. Entre los existentes destacar, por ejemplo, el uso de empaquetado sintáctico para las ideas [18] evaluando así el sentimiento implícito que presenta el texto o, por otro lado, la construcción de una base de conocimiento que mezcla el conocimiento afectivo con el sentido común [19].

2.2.4. Recursos para la detección de emociones

Los recursos existentes en el campo del procesamiento de lenguaje natural para la detección de emociones, como se ha comentado anteriormente, en comparación con el análisis de sentimiento, son escasos. Al tratarse de un campo relativamente nuevo dentro del procesamiento del lenguaje natural existe una falta de datos anotados disponibles. Las principales fuentes existentes para la anotación de datos son las siguientes:

- **Datos etiquetados:** se tratan de *datasets* de textos etiquetados con la emoción correspondiente. La existencia de un conjunto de datos anotados estándar, libre y generalizado supone un gran avance para el entrenamiento de cualquier tarea de clasificación de texto. En el campo de la detección de emociones no existe dicho estándar y, en general, es escaso el número de *datasets* etiquetados, pero cabe destacar la existencia de algunos recursos de interés. Uno de los principales *datasets* utilizados en el análisis de emociones es ISEAR [20], proporcionado por el *Swiss Center for Affective Sciences* (SCA), que se trata de un conjunto de 7.600 respuestas de diferentes personas sobre distintas situaciones que le han producido las emociones alegría, miedo, ira, tristeza, aversión, vergüenza o culpa, y cuál ha sido su reacción. Otro recurso a destacar es la base de conocimiento EmotiNet [21], que presenta otro punto de vista que consiste en dejar de actuar a nivel de palabra e interpretar la situación presentada en el texto, por ello la base de conocimiento está compuesta por cadenas de acción etiquetadas con la correspondiente emoción. Sin embargo la aplicación de EmotiNet en el análisis de emociones no resulta viable debido a su pequeño tamaño y a la estructura de la información planteada (solo cuatro tuplas formadas por actor, acción, objeto y emoción). Además, debido a la falta de *datasets*, también se utilizan recursos como emoticonos o *hashtags* para el etiquetado de textos procedentes de redes sociales, como por ejemplo el estudio de Wood y Ruder [22] que presenta un conjunto de los principales *emojis* utilizados por los usuarios para expresar su estado emocional dentro de las 6 categorías básicas de Ekman.
- **Léxicos de emociones:** consisten en diccionarios de palabras o conceptos que tienen asociados etiquetas afectivas. Como ocurre en el caso anterior, el número de léxicos

para el análisis de emociones es muy inferior a los existentes para el análisis de sentimiento. Uno de los léxicos más populares para la detección de emociones es WordNet-Affect [23], se trata de un léxico en inglés que asocia etiquetas de emociones a los *synsets* de Wordnet que representan conceptos afectivos. Este léxico contiene 2.874 *synsets* y 4787 palabras. Otro léxico a destacar es LIWC lexicón [24] que consiste en un diccionario compuesto de 2.300 palabras etiquetadas con polaridad (positiva o negativa) y categoría de emociones. Por otro lado, la existencia de recursos léxicos de emociones en español es muy reducido lo que supone una gran problemática para el análisis de textos en dicha lengua y la traducción automática de recursos en otros idiomas a este no ofrece resultados satisfactorios. Uno de los recursos en español existentes es Sentisense [25], un léxico afectivo basado en el concepto, que consiste en un conjunto de *synsets* de Wordnet etiquetados con una categoría emocional y que emplea los algoritmos de desambiguación propios de dicho léxico para solventar los problemas de ambigüedad. Otro léxico en español a destacar es Spanish Emotional Lexicon (SEL) [26] que contiene 2.036 palabras asociadas a un factor de probabilidad de la afectividad (PFA) correspondiente a al menos una de las 6 emociones básicas de Ekman.

- **Word Embeddings:** se trata de un conjunto de técnicas basadas en el modelado semántico distribuido, donde las palabras son representadas como vectores en un espacio n-dimensional de forma que la distancia entre vectores corresponda a la similitud semántica entre las palabras representadas. Este tipo de modelado ha proporcionado resultados satisfactorios en campos como la traducción automática [27] y el reconocimiento de entidades [28]. Su aplicación en el análisis de emociones podría mejorar el desempeño de los resultados, sin embargo solo se trata de un recurso potencial que carece de un desarrollo relevante.

2.3. Metodologías para la detección de emociones

Entre las metodologías utilizadas en la detección de emociones en textos, en concreto de textos procedentes de redes sociales, se pueden diferenciar 3 enfoques [29]: basado en palabras claves o léxicos, basado en aprendizaje automático e híbrido.

El método basado en palabras clave (*Keyword-based method*) se trata de un enfoque sencillo e intuitivo donde se pretende encontrar patrones en el texto analizado que encajen con palabras clave asociadas a emociones. Para ello es necesario dividir el texto a analizar en palabras y determinar la parte del discurso de cada una de ellas, identificando si se tratan de adjetivos, nombre, verbos o adverbios, que suponen las categorías gramaticales con mayor probabilidad de carga emocional. Una vez realizado dicho etiquetado, se comprueba dentro del conjunto de palabras etiquetadas la existencia de *keywords* asociadas a emociones.

La emoción que encaje con la frase analizada será la emoción asignada a dicha frase. Sin embargo, en el caso de presencia de distintas emociones se deben emplear mecanismos para la resolución de dicho conflicto. Por ejemplo, asociar un factor de probabilidad de emoción a cada *keyword* y escoger como emoción detectada aquella con mayor puntuación. Otra solución encontrada es la de escoger como emoción primaria aquella que ha sido encontrada en primer lugar.

El listado de palabras clave es construido por cada investigador mediante la selección de palabras asociadas a emociones y, en algunos casos, se utilizan diccionarios de sinónimos y antónimos para ampliar dicho listado

Por otro lado, el método basado en léxico (*Lexicon-based method*) es similar al anterior sin embargo utiliza léxicos de emociones en lugar de un listado de palabras clave para la detección de emociones. Algunos de los diferentes recursos existentes de este tipo han sido presentados en el subapartado 2.2.4 Recursos para la detección de emociones.

Entre estos dos métodos, principalmente destacan aquellos trabajos basados en léxicos de emociones. Por ejemplo, Badugu y Suhasini [30] proponen un sistema con un enfoque basado en reglas que clasifica *tweets* dentro de un conjunto de categorías emocionales. Dicha clasificación se basa en el modelo *Russell's Circumplex Model of Affect* [12], explicado en el subapartado 2.1 Representación de las emociones, determinando cuatro categorías emocionales: *Happy – Active Class*, *Happy – Inactive Class*, *Unhappy – Active Class* y *Unhappy – Inactive Class*. El léxico que han empleado para la base de conocimiento es el léxico Sentiwordnet 3.0 [31], que a pesar de tratarse de un diccionario de palabras para el

análisis de sentimiento, se han utilizado las puntuaciones de polaridad asociadas a cada palabra del léxico para la clasificación de las palabras en base a las 4 categorías mencionadas.

Las principales desventajas de este enfoque es la ambigüedad de las palabras, debido a que pueden presentar diferentes significados dependiendo del contexto, la incapacidad de detectar emoción en aquellos textos que no contengan las palabras clave, pese a que si expresen alguna emoción, y, por último, la falta de información lingüística, ya que no se tiene en cuenta ciertas estructuras sintácticas y semánticas que pueden influir en la emoción expresada.

Por otro lado, el otro enfoque empleado para la detección de emociones en texto y que goza de mayor popularidad es el uso de técnicas de aprendizaje automático, tanto supervisados como no supervisados. Dicho método permite la clasificación automática de textos a partir del diseño de un modelo para entrenar el clasificador a través de un número significativo de ejemplos que conformen los datos de entrenamiento. De esta forma se consigue clasificar de forma automática el resto de los datos (aprendiendo mediante un algoritmo que permite la identificación de comportamientos y/o características).

En el caso de los modelos supervisados, el conjunto de entrenamiento lo conforman datos etiquetados con emociones y los algoritmos de clasificación más utilizados son *Naive Bayes*, *Support Vector Machine* (SVM), los árboles de decisión y, en el campo específico de *Deep Learning*, las redes de neuronas LSTM por su capacidad de tener memoria para recoger las relaciones existentes entre los diferentes elementos del texto. Por otro lado, en los modelos no supervisados, los datos de entrenamiento no se encuentran etiquetados sino que el clasificador parte de un conjunto de palabras semilla asociadas a cada emoción que son referenciadas de forma cruzada con los datos de entrenamiento para así determinar la emoción de estos. A partir de dichos datos se entrena el clasificador no supervisado para determinar la emoción de los datos de prueba. En general, los resultados obtenidos por los modelos supervisados proporcionan una mayor precisión.

Un estudio a destacar dentro de este enfoque es el de Shah, Reyadh, Shaaf, Ahmed y Sithilen [32] que han trabajado en un sistema de detección de emociones en *tweets* sobre el *dataset* AIT-2018 [33]. El modelo propuesto se basa en los léxicos WordNet-Affect y EmoSentNet

[34] con la utilización de clasificadores supervisados (*Naïve Bayesian*, Árboles de decisión y Máquinas de Vectores Soporte). En este estudio la mayor precisión alcanzada ha sido del 88.23%, siendo utilizado para el entrenamiento el léxico EmoSenticNet y el algoritmo de Máquinas de Vectores Soporte.

Para el caso de detección de emociones en un corpus en español, siguiendo esta metodología, Blázquez, Berlanga y Molina [35] han realizado un sistema de inferencia de emociones de textos de Twitter en español siguiendo un enfoque supervisado, donde se han utilizado *hashtags* asociados a las seis emociones básicas de Ekman para la construcción del *dataset* de entrenamiento. El entrenamiento que han realizado parte de características con tamaño $n=1$ (unigrams) y emplea el algoritmo *Multinomial Naive Bayes* (MNB) para la clasificación de las emociones del corpus. La decisión de $n=1$ se realizó tras la prueba con varios tamaños (n igual a 1, 2, 3 y 1,2), siendo dicho tamaño el que obtuvo mayor precisión, un 65 %.

Un de las problemáticas que presenta este enfoque es la necesidad, en forma de características (*features*), de palabras clave. Por tanto, las problemáticas que presentan los métodos *keyword-based* también podrían estar presente en este enfoque. Además, también existe la necesidad de trabajo con un número significativo de ejemplos que conformen el conjunto de entrenamiento, lo que supone una problemática ante la falta de recursos existente en el campo del análisis de emociones

Por último, también se encuentra el uso de un enfoque híbrido. Este tipo de método combina los dos enfoques anteriores pudiendo aprovechar los beneficios de ambos enfoques y logrando una mayor precisión. Como ejemplo de este tipo de enfoque, destaca el trabajo “Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models,” [36], en el que se describe un modelo híbrido que utiliza un léxico chino, un sistema de reglas para la extracción de características y semántica relacionadas con las emociones y un clasificador basado en aprendizaje automático que utiliza los algoritmos SVM, *Naïve Bayesian* y *Max Entropy*.

2.4. Tecnologías para la detección de emociones.

Hoy en día, existen en el mercado diferentes soluciones comerciales que llevan a cabo la tarea de análisis de sentimiento o emociones en texto, por lo que el estudio de las mismas resulta de interés a la hora de definir la solución del presente proyecto. Entre las soluciones comerciales, la presencia de herramientas basadas en técnicas de aprendizaje automático es significativa, a continuación se describen alguna de ellas.

En primer lugar, como solución comercial a destacar, la API IBM Watson Tone Analyzer [37] detecta tanto la emoción expresada en un texto escrito (alegría, tristeza, etc.) como el tono del lenguaje utilizado (analítico, confiado y tentativo). La herramienta utiliza un modelo de aprendizaje automático basado en el algoritmo *Support Vector Machine* (SVM) aplicado a conversaciones para la predicción de nuevos textos. La respuesta ofrecida por IBM Watson Tone Analyzer es un conjunto de puntuaciones de las diferentes categorías de tonalidad del texto. Este tipo de software es de gran utilidad para compañías que quieran conocer el tono de las comunicaciones de sus clientes, a través de las redes sociales de la compañía, para así poder responderles adecuadamente y mejorar las conversaciones establecidas con estos.

En segundo lugar, la API ParallelDots Emotion Analysis [38] detecta la emoción de cualquier tipo de texto procedente de redes sociales, blogs o foros, estableciendo la probabilidad de pertenencia a las siguientes categorías: feliz, enfado, excitado, triste, miedo y aburrido. ParallelDots se basa en algoritmos de *Deep Learning* para la extracción de características y el clasificador se ha entrenado utilizando un algoritmo *Convolutional Neural Networks* (Covnets) mediante un *dataset* propio.

Por otro lado, con un enfoque basado en patrones de emociones, se encuentra la API MoodPatrol [39]. Se trata de una herramienta para desarrolladores que permite el análisis de textos y la detección de emociones. La clasificación utilizada se basa en el modelo de las 8 emociones básicas de Plutchik's, pudiendo detectar estas emociones básicas y una amplia variedad de emociones de mayor sutileza. Para ello, utiliza herramientas sofisticadas para el análisis de textos que permiten la identificación de patrones, en este caso asociados a emociones. Esta herramienta puede ser utilizada en la creación de aplicaciones o programas que busquen conocer cómo se sienten sus usuarios.

3. Análisis del sistema

En el presente capítulo, tras el estudio del estado del arte, se especificará la solución propuesta a la problemática planteada. Además, se especifican los requisitos a cumplir por sistema y el diseño de casos de uso que describen las acciones de interacción del usuario con el sistema y el comportamiento de este.

3.1. Análisis del problema.

A continuación, se define el problema a tratar y la solución escogida, justificando su elección frente otras posibles alternativas.

3.1.1. Definición del problema de ingeniería

La principal problemática que presenta el trabajo aquí expuesto es la obtención de conocimiento que sirva de soporte a la decisión mediante técnicas de análisis de fragmentos de textos de corta extensión en lenguaje natural (*tweets*). Dichos textos se obtienen a través de la plataforma existente SMART-RTVE, que permite la extracción desde redes sociales de una gran cantidad de recursos, en particular de la red social Twitter.

En concreto, el conocimiento a extraer se trata de la inferencia de emociones producto de los eventos deportivos a partir del texto generado en redes sociales por cuentas oficiales de los equipos, seguidores y medios de comunicación locales o nacionales asociados a dichos eventos. La salida obtenida producto del análisis será la emoción inferida para cada equipo y jornada y los *tweets* asociados a cada emoción extraída.

Por tanto, el problema planteado es el de la categorización de texto dentro del modelo clasificación de emociones escogido. En consecuencia, tras el estudio realizado en el estado del arte de los diferentes métodos, recursos y tecnología existentes en el campo del análisis de emociones, se define en el siguiente apartado la solución escogida.

3.1.2. Definición de la solución propuesta

La solución propuesta es la de un sistema basado en conocimiento que emplea un enfoque que utiliza palabras clave y léxicos para la detección de emociones en texto. Además de por

ser el método más sencillo e intuitivo, se ha escogido dicho enfoque frente al basado en aprendizaje, el cual goza de popularidad en el campo de la detección de emociones y ofrece resultados satisfactorios, debido a que su aplicación en este proyecto presenta ciertos inconvenientes.

En primer lugar, los resultados satisfactorios en los sistemas de aprendizaje aplicados al campo del análisis de emociones se producen principalmente en los sistemas supervisados, por lo que, debido a la escasez de *datasets* de entrenamiento en español existentes en este campo y la especificidad del dominio (eventos deportivos, en concreto partidos de fútbol de segunda división), se requeriría de una etiquetado emocional previo principalmente manual de una gran cantidad de dichos recursos, tarea costosa y que se debería repetir si se quisiese utilizar para otros dominios, tanto deportivos como de otra temática. En segundo lugar, si se decidiese utilizar un enfoque no supervisado, debido a la necesidad del uso de palabras clave en forma de características, las problemáticas presentes en los métodos *keyword-based* también afectarían a este enfoque. Por tanto, debido a la falta de recursos y la complejidad en la aplicación de aprendizaje automático en este proyecto se ha descartado el uso de dicho método. Se ha preferido un enfoque basado en el procesamiento del lenguaje natural y en una base de conocimiento a partir del cual aplicar reglas o algoritmos para la inferencia de emociones en el texto.

Por otro lado, cabe destacar que el uso de un enfoque híbrido resultaría altamente satisfactorio en cuanto resultados pero dados los plazos del proyecto no resulta abarcable. Sin embargo, se tiene presente para posibles trabajos futuros.

En cuanto al conocimiento en el que se basará el sistema, este estará conformado por el léxico SEL, que se trata de uno de los pocos léxicos de emociones en español disponibles, y el conjunto de *emojis* clasificados por el estudio de Wood y Ruder, ambos basados en el modelo de las 6 emociones básicas de Ekman. Se ha escogido el léxico SEL frente al léxico Sentisense, la otra alternativa de léxico en español encontrada, debido a que este último se trata de un léxico basado en concepto, en lugar de en términos, por lo que su inclusión en la herramienta utilizada de gestión del conocimiento basada en ontologías no es compatible.

Además, cabe destacar que en la solución adoptada se emplea la herramienta Knowledge Manager, que permite tanto la gestión del conocimiento en el que se basa el sistema como la aplicación de las técnicas de procesamiento del lenguaje natural necesarias para el análisis de texto.

Por último, se debe mencionar que la obtención de información relativa a los eventos deportivos analizados se realizará a través de un proveedor externo de información deportiva, el cual cuenta con una relación contractual con RTVE, y el resultado del análisis será proporcionado para su visualización por parte del usuario mediante interfaz gráfica.

3.2. Especificación de requisitos

En el siguiente apartado se recoge la especificación de los requisitos que debe cumplir el sistema. A continuación, se especifica el formato con el que se presenta cada requisito junto con la definición de los campos que se deben recoger:

ID	
Nombre	
Versión	
Prioridad	
Necesidad	
Fuente	
Estado	
Descripción	

Tabla 1. Ejemplo tabla de requisitos

La definición de los campos es la siguiente:

- **ID:** identificador único del requisito. Se trata de un código con el siguiente formato: RF/RFN – XY.
 - RF/RNF: indicador de tipo de requisito.
 - RF: requisito funcional.
 - RNF: requisito no funcional.
 - XY: código numérico, siendo X e Y números de 0 a 9.

- **Nombre:** nombre descriptivo del requisito.
- **Versión:** numeración para el control de cambios.
- **Prioridad:** nivel de urgencia de desarrollo del requisito. Los valores que contempla son alta, media o baja.
- **Necesidad:** nivel de importancia/obligatoriedad del requisito para el desarrollo del sistema. Los valores que contempla son esencial, deseado, opcional.
- **Fuente:** origen del requisito
- **Estado:** etapa del ciclo de vida del requisito en el que este se encuentra. Los valores contemplados son propuesto, aprobado, rechazado, implementado, validado
- **Descripción:** definición detallada del requisito.

A continuación, siguiendo el formato establecido y en base a los objetivos definidos anteriormente, se presentan una serie de requisitos distinguiendo dos grupos: requisitos funcionales y requisitos no funcionales.

3.2.1. Requisitos funcionales

Los requisitos funcionales recogen todas aquellas funcionalidades o capacidades que debe ofrecer el sistema al usuario para cumplir los objetivos definidos.

Contexto objetivo

ID	RF-01
Nombre	Extracción de contexto objetivo.
Versión	5.1
Prioridad	Alta
Necesidad	Esencial
Fuente	RTVE
Estado	Validado
Descripción	Para cada jornada, el sistema consumirá el servicio proporcionado por el servidor externo de información deportiva, obteniendo las noticias y metadatos referentes al grupo de interés de estudio (deporte fútbol, equipos de segunda B grupo IV).

Tabla 2. RF-01 Extracción de contexto objetivo

ID	RF-02
Nombre	Almacenamiento de contexto objetivo.
Versión	5.0
Prioridad	Media
Necesidad	Deseado
Fuente	Tutor
Estado	Validado
Descripción	El sistema almacenará la noticias y metadatos obtenidos del servicio externo proveedor de la información deportiva de interés.

Tabla 3. RF-02 Almacenamiento de contexto objetivo.

ID	RF-03
Nombre	Consulta de contexto objetivo.
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	<p>El sistema consultará los metadatos proporcionados por el servicio externo proveedor de información deportiva asociados a un equipo y jornada para obtener la información relevante del contexto objetivo:</p> <ul style="list-style-type: none"> • Equipo local • Equipo visitante • Fecha y hora del partido • Resultado

Tabla 4. RF-03 Consulta de contexto objetivo

Obtención y filtrado de recursos

ID	RF-04
Nombre	Monitorización de cuentas de usuarios en redes sociales.
Versión	5.2
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema monitorizará, mediante la plataforma de SMART-RTVE, las cuentas y <i>hashtags</i> de la red social Twitter cuyas publicaciones hagan referencia a los equipos de fútbol de Segunda B grupo IV para la obtención de dichos recursos.

Tabla 5. RF-04 Monitorización de cuentas de usuarios en redes sociales.

ID	RF-05
Nombre	Filtrado de recursos
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	RTVE
Estado	Validado
Descripción	El sistema filtrará los recursos de Twitter en base al equipo seleccionado por el usuario (equipos de la categoría 2B de Castilla La Mancha) y jornada.

Tabla 6. RF-05 Filtrado de recursos

ID	RF-06
Nombre	Eliminación de ruido en recursos
Versión	5.0
Prioridad	Baja
Necesidad	Deseado
Fuente	Tutor
Estado	Validado
Descripción	<p>El sistema proveerá de un mecanismo para la eliminación de aquellos recursos obtenidos que hagan referencia a información no relacionada:</p> <ul style="list-style-type: none"> • Otras categorías de equipo. • Eventos o sucesos no deportivos.

Tabla 7. RF-06 Eliminación de ruido en recursos

Detección de emociones

ID	RF-07
Nombre	Estrategias de clasificación de los recursos.
Versión	5.0
Prioridad	Baja
Necesidad	Opcional
Fuente	Tutor
Estado	Validado
Descripción	<p>En el sistema implementará 2 estrategias en el análisis de emociones de los recursos (contexto subjetivo) en función de la temporalidad a estudiar:</p> <ul style="list-style-type: none"> • Análisis contexto subjetivo semanal: análisis de los recursos creados desde la fecha de inicio de la semana en que transcurre la jornada hasta la hora de inicio del partido • Análisis contexto subjetivo de jornada: análisis de los recursos creados desde la hora de inicio del partido hasta las 3 horas siguientes.

Tabla 8. RF-07 Estrategias de clasificación de emociones.

ID	RF-08
Nombre	Clasificación emocional de los recursos.
Versión	5.1
Prioridad	Alta
Necesidad	Esencial
Fuente	RTVE
Estado	Validado
Descripción	El sistema clasificará automáticamente el contenido de los recursos para ambas estrategias en las siguientes categorías de emociones: tristeza, alegría, ira, miedo, aversión y sorpresa.

Tabla 9. RF-08 Clasificación emocional de los recursos

ID	RF-09
Nombre	Construcción de ontología.
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema utilizará la ontología definida del dominio para la detección de emociones y filtrado de ruido en los recursos.

Tabla 10. RF-09 Construcción de ontología

ID	RF-10
Nombre	Determinación de la emoción global en función del contexto objetivo
Versión	5.1
Prioridad	Baja
Necesidad	Opcional
Fuente	Tutor
Estado	Validado
Descripción	El sistema regulará la emoción global inferida del contexto subjetivo en función del resultado de la jornada (información extraída del contexto objetivo).

Tabla 11. RF-10 Determinación de la emoción global en función del contexto objetivo

ID	RF-11
Nombre	Servicio web de consulta
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	RTVE
Estado	Validado
Descripción	El sistema ofrecerá un servicio web de consulta de la emoción inferida en base a una jornada y dos equipos contrincantes.

Tabla 12. RF-11 Servicio web de consulta

ID	RF-12
Nombre	Resultado servicio web
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	RTVE
Estado	Validado
Descripción	El sistema ofrecerá como respuesta del servicio web la emoción inferida y los recursos asociados a dicha emoción una vez realizado el proceso de análisis de emociones.

Tabla 13. RF-12 Resultado servicio web

ID	RF-13
Nombre	Almacenamiento de observaciones
Versión	5.2
Prioridad	Media
Necesidad	Deseado
Fuente	Tutor
Estado	Validado
Descripción	El sistema almacenará el resultado del análisis de emociones para cada equipo y jornada especificando la siguiente información: emoción inferida, texto y porcentaje de <i>tweets</i> asociados a dicha emoción.

Tabla 14. RF-13 Almacenamiento de observaciones

Selección de parámetros y visualización de resultado.

ID	RF-14
Nombre	Selección de equipo
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	RTVE
Estado	Validado
Descripción	El sistema permitirá al usuario mediante interfaz seleccionar el equipo a analizar y se asignará de forma automática la última jornada.

Tabla 15. RF-14 Selección de equipo

ID	RF-15
Nombre	Selección de jornada
Versión	5.0
Prioridad	Baja
Necesidad	Opcional
Fuente	Tutor
Estado	Validado
Descripción	El sistema permitirá al usuario mediante interfaz, una vez seleccionado el equipo, seleccionar el número de jornada a analizar (pudiendo ser la última jornada o anteriores).

Tabla 16. RF-15 Selección de jornada

ID	RF-16
Nombre	Visualización de la emoción inferida
Versión	5.0
Prioridad	Media
Necesidad	Esencial
Fuente	RTVE
Estado	Validado
Descripción	El sistema ofrecerá la visualización el resultado del análisis (emoción y recursos asociados) mediante interfaz gráfica para las diferentes estrategias definidas.

Tabla 17. RF-16 Visualización de la emoción inferida

3.2.2. Requisitos no funcionales

Los requisitos no funcionales recogen aquellas restricciones que se deben seguir en el diseño y/o en la implementación, es decir, aquellas cualidades o propiedades que debe tener el sistema.

Rendimiento

ID	RNF-01
Nombre	Tiempo de respuesta del servicio web.
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema tendrá un tiempo de respuesta del servicio web de consulta de la emoción inferida no superior a 5 segundos.

Tabla 18. RNF-01 Tiempo de respuesta servicio.

ID	RNF-02
Nombre	Tiempo de carga de resultados en interfaz gráfica
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema tendrá un tiempo de carga de los resultados en interfaz gráfica no superior a 20 segundos.

Tabla 19. RNF-02 Tiempo de carga de resultados en Interfaz.

Entorno

ID	RNF-03
Nombre	Integración del sistema.
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema se deberá integrar en el entorno tecnológico de RTVE.

Tabla 20. RNF-03 Integración del sistema.

ID	RNF-04
Nombre	Resultado de análisis en web.
Versión	5.0
Prioridad	Media
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema visualizará el resultado del análisis, emoción inferida y <i>tweets</i> relacionados desde la web de SMART-RTVE

Tabla 21. RNF-04 Resultado de análisis en web.

ID	RNF-05
Nombre	Integración con servicio web proveedor de información deportiva.
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	RTVE
Estado	Validado
Descripción	El sistema se integrará con el servicio externo de información deportiva provisto por RTVE.

Tabla 22. RNF-05 Integración con servicio web proveedor de información deportiva.

Software

ID	RNF-06
Nombre	Framework de desarrollo
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema utilizará para su desarrollo el <i>framework</i> ASP.NET empleando el lenguaje de programación C#.

Tabla 23. RNF-06 Framework de desarrollo.

ID	RNF-07
Nombre	Tipo de servicio web.
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema de detección de emociones ofrecerá la emoción inferida y los tweets asociados mediante un servicio web de tipo REST.

Tabla 24. RNF-07 Tipo de servicio web.

Datos

ID	RNF-08
Nombre	Base de datos ontología.
Versión	5.0
Prioridad	Media
Necesidad	Deseable
Fuente	Tutor
Estado	Validado
Descripción	El formato de almacenamiento de la ontología será en MySQL para su gestión con la herramienta Knowledge Manager.

Tabla 25. RNF-08 Base de datos ontología.

ID	RNF-09
Nombre	Base de datos observaciones.
Versión	5.0
Prioridad	Media
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema empleará la base de datos ArangoDB para el almacenamiento de las observaciones sobre la inferencia de emociones.

Tabla 26. RNF-09 Base de datos observaciones.

ID	RNF-10
Nombre	Fuente de información deportiva.
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema sólo integrará información deportiva proveniente del servicio externo provisto por RTVE.

Tabla 27. RNF-10 Fuente de información deportiva.

ID	RNF-11
Nombre	Formato datos información deportiva.
Versión	5.0
Prioridad	Alta
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema recogerá los datos de información deportiva del servicio externo provisto por RTVE en formato 'JSON'.

Tabla 28. RNF-11 Formato datos información deportiva.

ID	RNF-12
Nombre	Base de datos información deportiva.
Versión	5.1
Prioridad	Media
Necesidad	Deseable
Fuente	Tutor
Estado	Validado
Descripción	El sistema empleará la base de datos ArangoDB para el almacenamiento de la información deportiva recogida del servicio externo provisto por RTVE.

Tabla 29. RNF-12 Base de datos información deportiva.

ID	RNF-13
Nombre	Periodicidad de generación de observaciones.
Versión	5.1
Prioridad	Media
Necesidad	Esencial
Fuente	Tutor
Estado	Validado
Descripción	El sistema generará las observaciones resultado del análisis de emociones por cada jornada, correspondiendo generalmente a un periodo semanal.

Tabla 30. RNF-13 Periodicidad de generación de observaciones.

3.3. Diseño de casos de uso

Tras la especificación de requisitos , en el siguiente apartado se realiza el diseño de casos de uso que consiste en la descripción de las acciones de interacción del usuario con el sistema y el comportamiento de este.

En primer lugar se lleva a cabo la representación gráfica de los casos de uso mediante un diagrama de casos de usos que permite identificar los actores que interaccionan con el sistema y las acciones que llevan a cabo. Posteriormente, se realiza una descripción detallada de cada caso de uso.

3.3.1. Diagrama de casos de uso

El diagrama resultante de la representación de los casos de uso del sistema es el siguiente:

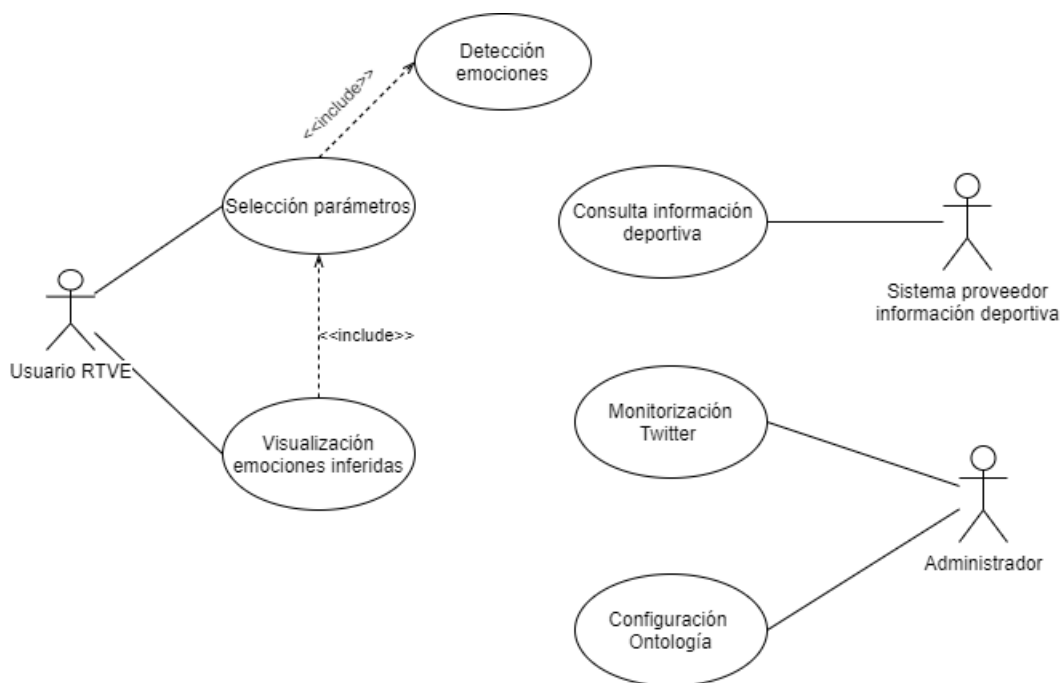


Figura 4. Diagrama de casos de uso

Como se puede observar en el diagrama, se identifican tres tipos de actores. En primer lugar, el Usuario de RTVE que interacciona con el sistema realizando la selección de equipo y jornada para la detección de emociones y, una vez realizado dicho análisis, visualiza el resultado del análisis de emociones. En segundo lugar, el sistema externo proveedor de información deportiva que interacciona con el sistema genérico mediante la consulta del servicio provisto para la obtención relativa al evento deportivo. En tercer lugar, el usuario Administrador que lleva a cabo las siguientes acciones en el sistema: monitorización de las cuentas para la obtención de recursos y configuración de la ontología para su posterior utilización en el proceso de detección de emociones.

Por último, cabe destacar la existencia de relaciones de dependencia de tipo <<include>> que indican que un caso de uso incluye el comportamiento del caso de uso del que depende. En el diagrama se observa que existe una dependencia del caso de uso ‘Visualización emociones inferidas’ hacia el caso de uso ‘Selección parámetros’ y, por otro lado, existe otra

dependencia del caso de uso ‘Selección parámetros’ hacia el caso de uso ‘Detección de emociones’.

3.3.2. Definición de casos de uso

A continuación, se describe cada uno de los casos de uso identificados en el diagrama anterior siguiendo el siguiente formato de tabla:

ID	
Nombre	
Actor	
Objetivo	
Precondiciones	
Postcondiciones	
Escenario básico	
Escenario alternativo	
Requisitos funcionales	

Tabla 31. Tabla ejemplo de descripción de casos de uso

La definición de los campos es la siguiente:

- **ID:** identificador único de caso de uso. Se trata de un código con el siguiente formato: CU- XY. Siendo X e Y números de 0 a 9.
- **Nombre:** título descriptivo de las acciones que recoge el caso de uso.
- **Actor:** agente que interacciona con el sistema con las acciones correspondientes al caso de uso. Los actores contemplados son: Usuario de RTVE , Administrador y Sistema proveedor de información deportiva.
- **Objetivo:** descripción del propósito del caso de uso.

- **Precondiciones:** condiciones que debe cumplir el sistema para que el caso de uso pueda llevarse a cabo.
- **Postcondiciones:** condiciones que presenta el sistema una vez se hayan producido las acciones correspondientes al caso de uso.
- **Escenario básico:** curso de eventos que se realizarán en el caso de uso.
- **Escenario alternativo:** conjunto de acciones alternativas que pueden ocurrir en el caso de uso.
- **Requisitos funcionales:** requisitos funcionales (especificados en 3.2 Especificación de requisitos) relacionados con el caso de uso.

ID	CU-01
Nombre	Selección parámetros
Actor	Usuario RTVE
Objetivo	Establecimiento de los equipos parámetro del posterior análisis de emociones
Precondiciones	El usuario de RTVE ha accedido con credenciales validas a la plataforma SMART-RTVE.
Postcondiciones	El sistema muestra por interfaz la selección realizada
Escenario básico	<ol style="list-style-type: none"> 1. El usuario RTVE accede a la plataforma web 2. El usuario RTVE selecciona en la interfaz el equipo sobre el que desea realizar el análisis 3. El sistema asigna de forma automática la última jornada. 4. El sistema consulta de base de datos el contexto objetivo para establecer los siguientes parámetros: equipo local, equipo visitante, resultado y fecha y hora del partido 5. El sistema llama al servicio web de consulta de detección de emociones con los parámetros establecidos (equipo local, equipo visitante, jornada, resultado, fecha y hora del partido).
Escenario alternativo	<ol style="list-style-type: none"> 1. El usuario RTVE accede a la plataforma web 2. El usuario RTVE selecciona en la interfaz el equipo sobre el que desea realizar el análisis 3. El usuario RTVE selecciona una jornada anterior. 4. El sistema consulta de base de datos el contexto objetivo para establecer los siguientes parámetros: equipo local, equipo visitante, resultado y fecha y hora del partido 5. El sistema llama al servicio web de consulta de detección de emociones con los parámetros establecidos (equipo local, equipo visitante, jornada, resultado, fecha y hora del partido).
Requisitos funcionales	RF-03, RF-11, RF-12, RF-14, RF-15

Tabla 32. CU-01 Selección parámetros

ID	CU-02
Nombre	Detección emociones
Actor	Usuario RTVE
Objetivo	Detección de emociones en los recursos asociados a un equipo y jornada
Precondiciones	El usuario RTVE ha seleccionado equipo y jornada a analizar
Postcondiciones	En base de datos se encuentra almacenado el resultado del análisis y dicho resultado se ha devuelto como respuesta del análisis
Escenario básico	<ol style="list-style-type: none"> 1. El sistema comprueba que para los parámetros establecidos en base a la selección del usuario RTVE no existen observaciones sobre análisis de emociones 2. El sistema obtienen de base de datos los recursos asociados a los parámetros establecidos en base a la selección del usuario RTVE. 3. El sistema realiza el análisis de emociones en base a las estrategias establecidas (contexto subjetivo semanal y jornada) 4. El sistema regula la intensidad de la emoción inferida global en base al resultado (contexto objetivo). 5. El sistema devuelve y almacena el resultado del análisis (emoción inferida y <i>tweets</i> asociados para cada equipo).
Escenario alternativo	<ol style="list-style-type: none"> 1. El sistema comprueba que para los parámetros establecidos en base a la selección del usuario RTVE si existen observaciones sobre análisis de emociones 2. El sistema consulta dichas observaciones sobre el resultado del análisis. 3. El sistema devuelve el resultado del análisis (emoción inferida y <i>tweets</i> asociados para cada equipo).
Requisitos funcionales	RF-05, RF-06, RF-07, RF-08, RF-10, RF-13

Tabla 33. CU-02 Detección emociones

ID	CU-03
Nombre	Visualización emoción inferida
Actor	Usuario RTVE
Objetivo	Visualización, a través de la interfaz, del resultado obtenido del análisis de emociones.
Precondiciones	El usuario RTVE ha seleccionado los parámetros y se ha llevado a cabo en análisis de emociones
Postcondiciones	La interfaz muestra la emoción inferida y los <i>tweets</i> asociados para cada equipo.
Escenario básico	<ol style="list-style-type: none"> 1. El usuario recibe la respuesta del servicio web de consulta de detección de emociones. 2. El usuario visualiza los datos de respuesta (emoción inferida y <i>tweets</i> asociados) para cada equipo (local y visitante)
Escenario alternativo	No aplica.
Requisitos funcionales	RF-12, RF-16

Tabla 34. CU-03 Visualización emoción inferida

ID	CU-04
Nombre	Consulta información deportiva
Actor	Sistema proveedor información deportiva
Objetivo	Obtención de la información relativa al contexto objetivo de la jornada.
Precondiciones	Se ha producido una nueva jornada.
Postcondiciones	En base de datos se encuentran almacenadas las noticias y metadatos proporcionados por el sistema proveedor de información deportiva.
Escenario básico	<ol style="list-style-type: none"> 1. El sistema de proveedor de información deportiva genera las noticias y metadatos para la última jornada producida. 2. El sistema consulta dicha información a través del servicio externo provisto. 3. El sistema almacena la información proporcionada.
Escenario alternativo	No aplica.
Requisitos funcionales	RF-01, RF-02

Tabla 35. CU-04 Consulta información deportiva

ID	CU-05
Nombre	Monitorización Twitter
Actor	Administrador
Objetivo	Monitorización de cuentas y <i>hashtags</i> de Twitter para la obtención de recursos a analizar.
Precondiciones	El Administrador del sistema ha accedido a la plataforma SMART-RTVE
Postcondiciones	El contenido de las cuentas y <i>hashtags</i> monitorizados se encuentra almacenado en base de datos generando recursos.
Escenario básico	<ol style="list-style-type: none"> 1. El Administrador del sistema establece las cuentas y <i>hashtags</i> de Twitter a monitorizar. 2. La plataforma SMART-RTVE monitoriza dichas fuentes de forma periódica. 3. La plataforma SMART-RTVE almacena en base de datos las publicaciones asociadas a las cuentas y <i>hashtags</i> generando así recursos.
Escenario alternativo	No aplica.
Requisitos funcionales	RF-04.

Tabla 36. CU-05 Monitorización Twitter

ID	CU-06
Nombre	Configuración ontología
Actor	Administrador
Objetivo	Configuración de la ontología que supone la base de conocimiento para el análisis de emociones y filtrado de recursos no relevantes.
Precondiciones	El Administrador del sistema cuenta con licencia de la herramienta Knowledge Manager
Postcondiciones	La ontología para la detección de emociones se encuentra definida en la base de datos MySQL.
Escenario básico	<ol style="list-style-type: none"> 1. El Administrador del sistema, mediante la herramienta Knowledge Manager, define la terminología y los patrones a utilizar para la detección de emociones. 2. El Administrador del sistema, mediante la herramienta Knowledge Manager, define la terminología y los patrones a utilizar para la eliminación de ruido en los recursos a analizar. 3. La herramienta Knowledge Manager almacena los cambio de la ontología en base de datos. 4. El sistema utiliza la ontología definida para la detección de emociones y filtración de ruido en los recursos.
Escenario alternativo	No aplica.
Requisitos funcionales	RF-06, RF-08, RF-09.

Tabla 37. CU-06 Configuración ontología

4. Diseño del Sistema

En el presente capítulo se explica detalladamente el diseño de la solución propuesta basado en el análisis del sistema realizado en el punto anterior. Se describe tanto la infraestructura tecnológica empleada para el desarrollo del sistema, el diseño de la arquitectura de la solución y el diseño del proceso de análisis y la ontología empleada para la detección de emociones.

4.1. Infraestructura tecnológica

En este apartado, se presentan las diferentes herramientas software y *frameworks* utilizados en el desarrollo del sistema propuesto. Cabe destacar que para la coordinación y colaboración eficiente con las otras partes que conforman la totalidad del proyecto Social Media News (presentado en el subapartado 1.1.1 Proyecto Social Media News) se ha empleado un modelo basado en software libre y “Devops” (*continuous integration and delivery*). La infraestructura empleada para soportar el modelo definido es la ofrecida por Github [40], plataforma para el desarrollo colaborativo que utiliza un sistema de control de versiones. Los proyectos alojados en Github son de acceso libre, sin embargo, se ha utilizado un repositorio privado debido al acuerdo de colaboración con RTVE.

4.1.1. Microsoft Visual Studio

Microsoft Visual Studio [41] se trata de un entorno de desarrollo integrado (IDE), disponible para varios sistemas operativo (Windows, MacOS, Linux, etc.), que ofrece un conjunto de herramientas para el diseño, implementación y pruebas en el desarrollo de aplicaciones. El número de lenguajes de programación soportados es elevado, incluyendo aquellos con mayor popularidad tales como C#, C ++, .NET, Python, PHP o Java.

Entre las diferentes herramientas y componentes ofrecidos, Microsoft Visual Studio proporciona un entorno para el desarrollo web con ASP.NET (.NET Framework) [42] que se trata de un *framework* para la creación de sitios y aplicaciones web.

En el presente proyecto se ha escogido el IDE Visual Studio Enterprise 2019. Dentro de este, se ha utilizado el entorno ASP.NET MVC, para la creación de los servicios web REST

encargados de recibir las peticiones de los usuarios realizadas mediante interfaz y su traslado al sistema de análisis.

4.1.2. C#

El lenguaje de programación C# [43] se trata de un lenguaje orientado a objetos desarrollado por Microsoft y es uno de los lenguajes utilizados en el entorno ASP.NET. C# proviene de C/C++ e incluye un modelo de objetos con similitudes con el de Java.

Este lenguaje será utilizado en el sistema propuesto tanto para la implementación del servicio web como en el sistema de análisis de emociones. Esto se debe a su compatibilidad para desarrollo web ASP.NET y al hecho de que el resto de las partes del proyecto Social Media News y, en general, los proyectos de la cátedra RTVE-UC3M, utilizan dicho lenguaje, por lo que el uso de este supone una mayor integración y capacidad de reutilización del sistema desarrollado.

4.1.3. Angular

Angular [44] se trata de un *framework* creado por Google de código abierto, utilizado para el desarrollo de páginas y aplicaciones web de tipo SPA (*Single Page Application*). Una de sus principales finalidades es la de dar soporte a aquellas aplicaciones o sistemas que se basan en un navegador con capacidad Modelo Vista Controlador (MVC). Angular utiliza TypeScript como lenguaje de desarrollo que se trata de un superconjunto de Javascript y proporciona una extensión de HTML mediante el uso de directivas y atributos. Además, cuenta con una estructura basada en componentes que se tratan de un tipo de clase que cuenta con propiedades (variables) y métodos (funciones).

Angular, en concreto Angular 6, es utilizado en el sistema para el desarrollo de la web para la muestra final de resultados al usuario dentro de la plataforma existente SMART-RTVE. Los lenguajes empleados en dicho desarrollo son TypeScript, como es propio de angular, HTML y CSS. Además, se emplea la librería HighCharts [45] desarrollada en Javascript para la creación de gráficos interactivos en sitios o aplicaciones web que pueden ser importados a Angular 6 para su uso en la muestra del resultado del análisis de emociones.

4.1.4. ArangoDB

ArangoDB [46] es una base de datos no relacional multi-modelo que soporta los modelos clave-valor, documentos y grafos. El lenguaje de consulta para esta base de datos es AQL (*ArangoDB Query Language*), un lenguaje propio similar a SQL que permite la consulta entre los diferentes modelos soportados. La gestión de la base de datos se puede realizar mediante una interfaz web o a través del cliente de ArangoSH. Entre las principales ventajas que ofrece ArangoDB se encuentran el modelado flexible de datos y su alto rendimiento.

La plataforma SMART-RTVE, utilizada en el presente proyecto para la monitorización de redes sociales, emplea ArangoDB para el almacenamiento de los datos extraídos de Twitter. En la construcción de dicha plataforma se decidió utilizar ArangoDB por la flexibilidad ofrecida en cuanto a los diferentes modelos de datos soportados, su compatibilidad con diferentes sistemas operativos y el rendimiento ofrecido.

En el sistema de análisis de emociones propuesto consultará los recursos almacenados de Twitter desde la solución implementada mediante el cliente de ArangoDB para .NET versión 0.7.70.0, realizando dichas consultas en lenguaje AQL. Además, también se utilizará ArangoDB para el almacenamiento de las observaciones resultado del análisis para su integración con la plataforma ya existente.

4.1.5. MySQL

MySQL [47] es uno de los sistemas gestor de base de datos relacionales de mayor popularidad. En cuanto a la licencia para su uso presenta una licencia dual, por un lado una la licencia pública general que permite el uso gratuito por la comunidad y código abierto, y, por otro lado, licencia comercial propiedad de Oracle. Además, el motor MySQL se puede utilizar en los diferentes sistemas operativos Windows, Linux y Mac.

En el presente proyecto MySQL es utilizado como base de datos para el almacenamiento de la ontología creada por la herramienta de gestión del conocimiento Knowledge Manager (KM). MySQL es una de las bases de datos soportadas por dicha herramienta y se ha escogido esta frente a otras, como por ejemplo Microsoft Access, por la capacidad de soportar Unicode para el almacenamiento desde KM de *emojis* en la ontología. Una vez establecida desde KM

la conexión con el servidor de la base datos MySQL, toda la gestión de la misma se realizará desde la herramienta mencionada.

4.1.6. Knowledge Manager

Tal como se ha comentado anteriormente en el estado del arte (subapartado 2.2.2 Herramientas para el análisis de texto), Knowledge Manager es una herramienta de gestión del conocimiento propiedad de *The Reuse Company*. Entre las diferentes funcionalidades ofrecidas se encuentra la creación y gestión de ontologías y vocabulario, creación de modelos para la representación del conocimiento y creación y uso de patrones para el análisis de texto, entre otras.

En el sistema de análisis de emociones se ha decidido emplear la API ofrecida por KM, por un lado, para la creación del vocabulario compuesto por el léxico y el conjunto de *emojis* asociados a emociones y la gestión de este, y, por otro lado, para la creación de patrones e indexación de los recursos de entrada del sistema con la finalidad de emplear las técnicas de *pattern matching* que permiten la identificación de las emociones presentes en el texto.

El uso de esta herramienta resulta fundamental para llevar a cabo el enfoque basado en conocimiento escogido para desarrollar el sistema ya que permite llevar a cabo todas las tareas relacionadas con la gestión del conocimiento y procesamiento del lenguaje natural necesarias para el análisis.

4.2. Definición de la Arquitectura del sistema

Dada la arquitectura completa del sistema Social Media News, presentada en el capítulo 1, subapartado 1.1.1 Proyecto Social Media News, el componente de análisis de emociones propuesto se encuentra integrado en dicho sistema como muestra la siguiente figura.

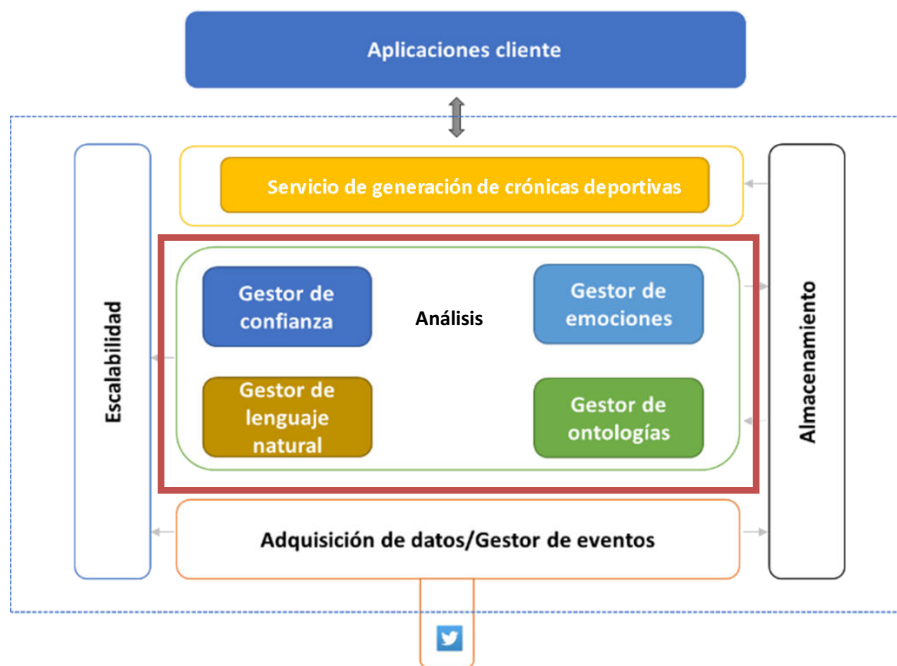


Figura 5. Arquitectura de alto nivel del sistema Social Media News (Módulo análisis)

Como se ha comentado anteriormente, el módulo de análisis de emociones propuesto pretende proporcionar al sistema de Social Media News la funcionalidad de identificación de emociones entre los textos generados en redes sociales (en concreto Twitter) relativos a los eventos deportivos de interés, para así conocer el estado emocional de los aficionados asociado a una jornada/partido disputado.

En los siguientes apartados, para una descripción de la arquitectura del módulo de análisis, se seguirá el modelo 4+1 [48] en el cual se establecen las siguientes vistas:

- **Vista lógica:** consiste en la definición del modelo conceptual que recoja los requerimientos funcionales del sistema de detección de emociones. La representación de dicho modelo se ha llevado a cabo mediante un diagrama de clases de UML, en el que se establece una serie de clases y las relaciones lógicas existentes entre las mismas.
- **Vista de desarrollo:** especifica los diferentes componentes que conforman el sistema de análisis de emociones y las dependencias existentes entre estos. Se ha empleado un diagrama de componentes para su representación.

- **Vista de casos:** muestra el comportamiento del sistema desde la perspectiva del usuario, describiendo la secuencia de interacciones entre los agentes externos y el sistema. Esta vista se encuentra especificada en el análisis del sistema, en el subapartado 3.3 Diseño de casos de uso.

4.2.1. Vista lógica

En este apartado se describe, en base a un método de diseño orientado a objetos, el modelo conceptual del sistema mediante un diagrama de clases de UML, donde el sistema se muestra como una serie de clases de objetos y sus relaciones lógicas: asociación, herencia, dependencia, etc.

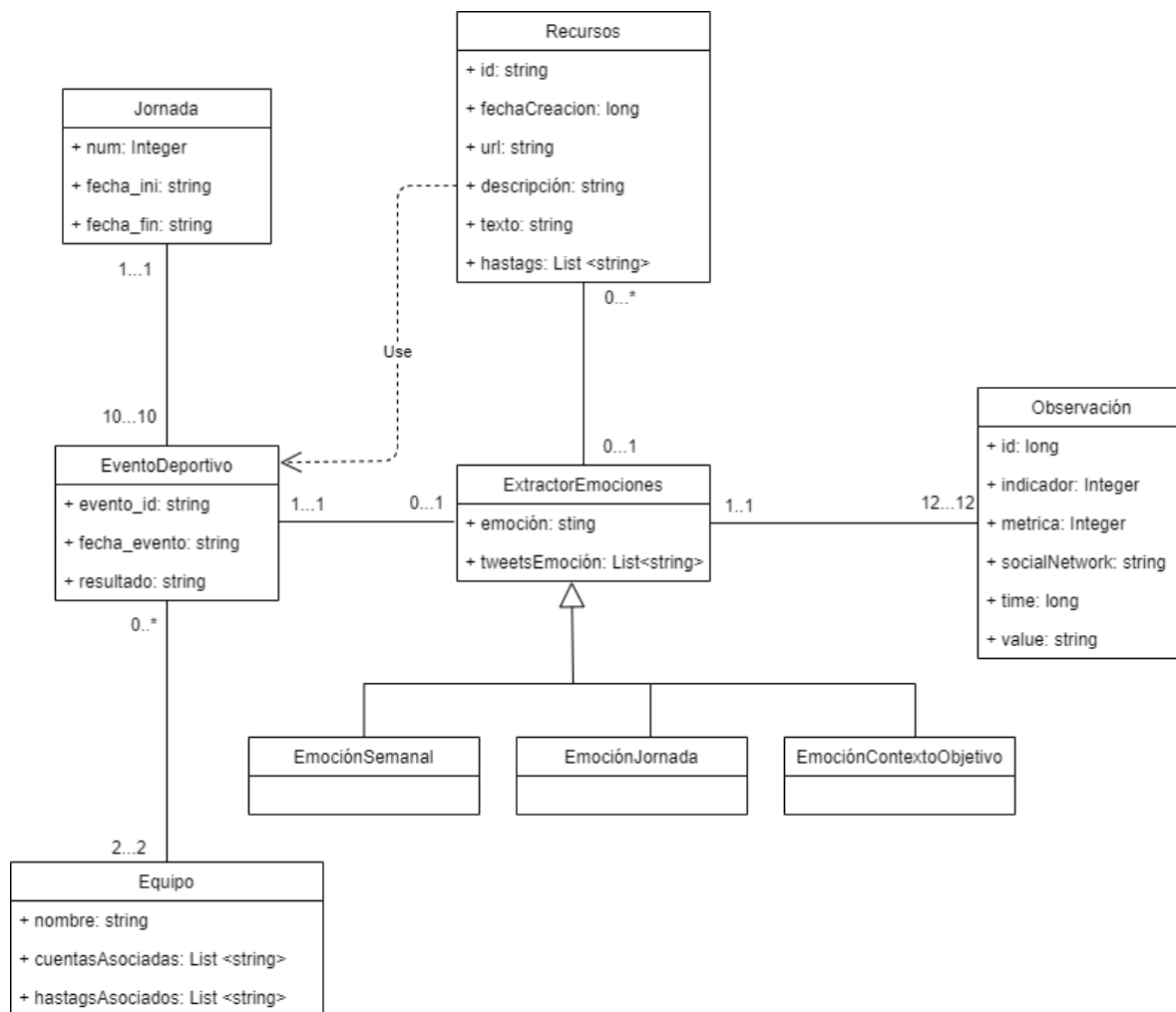


Figura 6. Diagrama de clases

A continuación se explica las diferentes clases que componen el modelo conceptual del sistema representado en el diagrama anterior, sus atributos así como la relación existente entre las mismas:

- **Jornada:** clase que representa la información relativa a la jornada a analizar. Esta clase presenta una relación de asociación con multiplicidad 10...10 con la clase 'EventoDeportivo' ya que una jornada siempre tendrá asociado 10 eventos deportivos, uno por cada 2 equipos que conforman la liga analizada. Los atributos propios de la clase son los siguientes:
 - Num: atributo de tipo *Integer* que indica el número de jornada.
 - Fecha_ini: atributo de tipo *string* que recoge la fecha de inicio de la jornada.
 - Fecha_fin: atributo de tipo *string* que contiene la fecha del final de la jornada.
- **Equipo:** clase que modeliza a los equipos objeto de análisis. Esta clase se relaciona mediante asociación con la clase 'EventoDeportivo' con multiplicidad 0...*, es decir, un equipo puede tener asociado en el sistema infinitud de eventos, tantos como transcurran en la liga analizada. Los atributos propios de esta clase son los siguientes:
 - Nombre: atributo de tipo *string* que contiene el nombre del equipo, dicho nombre es único para los equipos estudiados en el sistema.
 - CuentasAsociadas: listado de datos tipo *string* que recogen las cuentas de Twitter asociadas al equipo que permiten la consulta sus recursos.
 - HashtagsAsociados: listado de datos tipo *string* que recogen los *hashtags* de Twitter asociados al equipo que permiten la consulta sus recursos.
- **EventoDeportivo:** clase que representa la información relativa a los eventos deportivos objeto de estudio del sistema. Esta clase presenta relación de asociación con tres clases, 'Jornada', 'Equipo' y 'ExtractorEmociones'. En primer lugar, existe una asociación con multiplicidad 1..1 con 'Jornada', lo que indica que a un evento siempre le corresponde una única jornada. En segundo lugar, la relación con 'Equipo' presenta una multiplicidad 2..2, por tanto, el evento siempre tiene asociados 2 equipos. Por último, con la clase 'ExtractorEmociones' existe una relación de asociación 0...1, lo que implica que por cada evento como máximo se le asocia una

instancia del extractor de emociones. Los atributos propios de esta clase son los siguientes:

- Evento_id: atributo de tipo *string* que permite la identificación inequívoca del evento deportivo.
 - Fecha_evento: atributo de tipo *string* que almacena la fecha y hora del evento.
 - Resultado: atributo de tipo *string* que recoge el resultado numérico del evento.
- **ExtractorEmociones**: clase que modeliza las actividades de detección de emociones del sistema. Esta clase presenta relación de asociación con tres clases, 'EventoDeportivo', 'Recursos' y 'Observación'. En primer lugar, existe una relación de asociación con la clase 'EventoDeportivo' de multiplicidad 1..1 que indica que un proceso de análisis de emociones siempre tiene asociado 1 evento. En segundo lugar, esta clase presenta una relación de asociación de multiplicidad 0..* con la clase 'Recursos', lo que representa que por cada proceso de análisis se estudian infinitud de recursos, tantos como los que se publiquen para el evento analizado. Por último, existe una relación de asociación con la clase 'Observaciones' con multiplicidad 12..12 ya que para cada proceso de detección de emociones se almacenan 12 observaciones resultado del análisis, 2 observaciones por cada equipo y tipo de análisis (semanal, jornada y con contexto objetivo) . Los atributos propios de esta clase son los siguientes:
 - Emoción: atributo de tipo *string* que contiene la emoción extraída del análisis. Su valor solo puede corresponder a una de las siguientes categorías: Emoción no detectada, Alegría, Tristeza, Sorpresa, Ira, Aversión y Miedo.
 - TweetsEmocion: listado de datos de tipo *string* que contienen los identificadores de los *tweets* asociados al valor del atributo emoción.
 - **EmocionSemanal, EmocionJornada y EmocionContextoObjetivo**: clases hija o derivadas de la clase ExtractorEmociones cuya representación surge de la necesidad de representar las diferentes emociones a analizar en el sistema: análisis de la emoción semanal (EmociónSemanal), detección de la emoción correspondiente al transcurso del evento (EmocionJornada) y análisis de la emoción teniendo en cuenta el resultado del evento (EmocionContextoObjetivo).

- **Recursos:** clase que representa los *tweets* recogidos por el sistema para su análisis. Por un lado, esta clase presenta una relación de asociación 0...1 con la clase 'ExtractorEmociones' que implica que un recurso puede ser analizado como máximo por un proceso de detección de emociones asociado a un evento. Por otro lado, se establece una relación de dependencia con la clase 'EventoDeportivo' ya que requiere de la información contenida en dicha clase para la obtención de los recursos asociados a un evento. Los atributos propios de esta clase son los siguientes:
 - **Id:** atributo de tipo *string* que permite la identificación inequívoca del recurso.
 - **FechaCreación:** atributo de tipo *long* que almacena la fecha de creación del recurso en milisegundos.
 - **Url:** atributo de tipo *string* que contiene la url que permite la localización y acceso a la publicación en la web de Twitter.
 - **Descripción:** atributo de tipo *string* que recoge un texto descriptivo que indica la cuenta creadora del recurso y las cuentas que lo han retweetado.
 - **Texto:** atributo de tipo *string* que contiene el texto del *tweet*.
 - **Hashtags:** listados de datos tipo *string* que recogen los *hashtags* mencionados en el *tweet*.
- **Observación:** clase que representa las observaciones producto del análisis de emociones que serán almacenadas en base de datos y empleadas para la visualización de resultado. Esta clase presenta una relación de asociación de multiplicidad 1 con la clase 'ExtractorEmociones', lo que indica que una observación siempre está asociado a un proceso de análisis de emociones. Los atributos propios de esta clase son los siguientes:
 - **Id:** atributo de tipo *string* que permite la identificación inequívoca de la observación.
 - **Indicador:** atributo de tipo *Integer* que recoge tipo de análisis: 1 (jornada), 2 (semanal) y 3 (con contexto objetivo)
 - **Métrica:** atributo de tipo *Integer* que especifica el tipo de observación dentro del tipo de análisis indicado en el atributo Indicador. Los valores posibles son: 'valorIndicador'+ '01' (Emoción global) y 'valorIndicador'+ '02' (Porcentaje resultante de cada emoción estudiada).

- socialNetwork: atributo de tipo *string* que indica el nombre de la red social analizada. En el presente proyecto solo se analizarán recursos procedentes de Twitter pero se establece dicho atributo para un futuro análisis de otras redes sociales.
- value: atributo de tipo *string* que especifica el valor de la observación resultado del análisis de emociones.
- time: atributo de tipo *long* que recoge la fecha de creación de la observación en milisegundos.

4.2.2. Vista de Desarrollo

En este apartado, se explica la organización de los módulos software en el entorno de desarrollo del sistema y las dependencias existentes entre dichos componentes, siguiendo el patrón Modelo-Vista-Controlador (MVC).

El modelo arquitectónico MVC [49] se basa en la separación de los componentes en tres categorías, Modelo, Vista y Controlador, en función de su finalidad en el sistema:

- **Modelo:** tipo de componente responsable de la lógica de negocio de los datos de la aplicación. Se encarga de las tareas de acceso a base de datos y definición de las reglas de lógica de negocio a aplicar a los datos en base a la funcionalidad del sistema.
- **Vista:** se trata de la interfaz de usuario, por lo que es responsable de proporcionar al usuario la información ofrecida por sistema así como del manejo de los mecanismos de interacción con el mismo.
- **Controlador:** tipo de componente encargado de la lógica de control del sistema, actuando de intermediario entre la Vista y el Modelo. Es el responsable de capturar los eventos de entrada provenientes de la vista y gestionar dichos eventos según las reglas de control definidas para el acceso al modelo.

A continuación, se presenta la vista de desarrollo del sistema mediante un diagrama de componentes UML, como muestra la Figura 7.

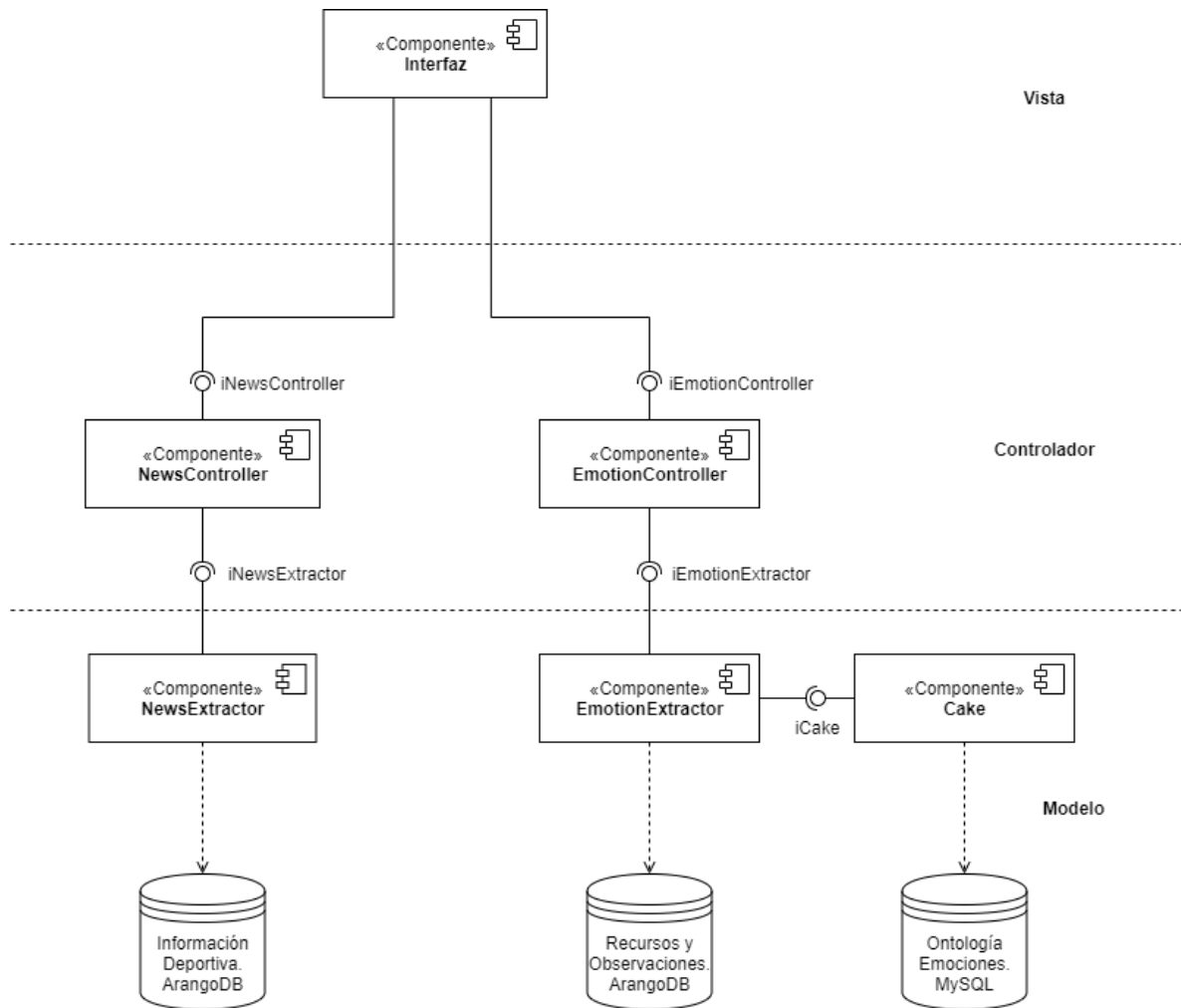


Figura 7. Diagrama de componentes

A continuación, se especifica la descripción de cada componente, indicando su funcionalidad, dependencias, entrada, salida y el contrato de operaciones de los métodos implementados:

Componente	Interfaz	
Propósito	Componente web que proporciona la interfaz de usuario que permite la interacción de este con el sistema, seleccionando los parámetros de entrada del mismo (equipo y jornada), y la visualización del resultado del análisis.	
Dependencias	iControlador	
Interfaz	iInterfaz	
Contrato de operaciones	Obtener_Informacion_Evento (string equipo, int jornada)	
	Descripción	Muestra la información relativa al evento deportivo asociado al equipo y jornada de entrada.
	Precondiciones	El usuario ha seleccionado un equipo y jornada mediante interfaz gráfica.
	Postcondiciones	<ul style="list-style-type: none"> La función devuelve un objeto con los siguientes campos relativos a la información del evento: <ul style="list-style-type: none"> Equipo local Equipo visitante Resultado Fecha y hora del evento La interfaz muestra la información sobre los parámetros seleccionados y equipo local y visitante
	Obtener_Emocion_Evento (string equipo_local, string equipo_visitante, int jornada, long fecha_evento, int resultado_evento)	
	Descripción	Muestra la emoción global obtenida para cada equipo seleccionado para los diferentes tipos de análisis establecidos.

	Precondiciones	Se debe haber llamado a la función <code>Obtener_Informacion_Evento()</code> para tener la información requerida como entrada de la función.
	Postcondiciones	La interfaz muestra gráficamente la emoción obtenida.
	Obtener_Gráfico_Emocion (string equipo, int jornada, string [] porcentajes)	
	Descripción	Genera un gráfico que recoge el porcentaje de <i>tweets</i> asociados a cada emoción para cada equipo, jornada y tipo de análisis.
	Precondiciones	Deben existir datos registrados sobre el porcentaje de <i>tweets</i> asociados a cada emoción para el equipo y jornada de entrada.
	Postcondiciones	Muestra mediante interfaz gráfica el gráfico generado.

Tabla 38. *Componente Interfaz*

Componente	NewsController	
Propósito	Componente encargado de recibir y gestionar aquellos eventos provenientes de la interfaz relativos a la funcionalidad de extracción de información del evento deportivo. Actúa como intermediario entre la Interfaz y el componente modelo NewsExtractor.	
Dependencias	iNewExtractor	
Interfaz	iNewsController	
Contrato de operaciones	get_Informacion_Evento (string equipo, int jornada)	
	Descripción	Servicio web para la obtención la información relativa al evento deportivo asociado al equipo y jornada de entrada.
	Precondiciones	Los argumentos deben cumplir lo siguientes requisitos: <ul style="list-style-type: none"> • Equipo no puede ser nulo • El valor de jornada debe ser un número de jornada válido
	Postcondiciones	La función devuelve un objeto con los siguientes campos: <ul style="list-style-type: none"> • Código del estado de la petición • Objeto que recoge la información del evento consultado.
	post_Ultimo_Evento ()	
	Descripción	Servicio web para la de consulta de los eventos de la última jornada a través del servicio externo proveedor de información deportiva y el almacenamiento de dicha información en base de datos
	Precondiciones	Debe estar disponible en el servicio proveedor de información deportiva la información sobre los eventos de la última jornada.

	Postcondiciones	La información de los eventos de la última jornada se encuentran almacenados en base de datos.
	Get_Ultima_Jornada ()	
	Descripción	Servicio web de consulta sobre el número de la última jornada sobre la que se tiene eventos deportivos almacenados en base de datos.
	Precondiciones	Deben existir datos registrados en base de datos sobre los eventos deportivos objeto de estudio.
	Postcondiciones	<p>La función devuelve un objeto con los siguientes campos:</p> <ul style="list-style-type: none"> • Código del estado de la petición • Número que indica la última jornada almacenada.

Tabla 39. Componente NewsController

Componente	EmotionController	
Propósito	Componente encargado de recibir y gestionar aquellos eventos provenientes de la interfaz relativos a la funcionalidad de detección de emociones. Actúa como intermediario entre la Interfaz y el componente modelo EmotionExtractor.	
Dependencias	iEmotionExtractor	
Interfaz	iEmotionController	
Contrato de operaciones	get_Emocion_Semanal (string equipo_local, string equipo_visitante, int jornada, long fecha_evento)	
	Descripción	Servicio web para la obtención de la emoción global extraída para cada equipo (local y visitante) en el contexto semanal junto a los <i>tweets</i> asociados a dicha emoción .
	Precondiciones	<ul style="list-style-type: none"> Los argumentos deben cumplir lo siguientes requisitos: <ul style="list-style-type: none"> Equipo local y visitante no pueden ser nulos El valor de jornada y fecha_evento debe ser mayor a 0. Se debe haber obtenido previamente la información relativa al evento deportivo.
	Postcondiciones	La función devuelve un objeto con los siguientes campos: <ul style="list-style-type: none"> Código del estado de la petición Listado de objetos que recoge la información sobre la detección de emociones para cada equipo (Emoción extraída y Listado de <i>tweets</i> asociados)
	get_Emocion_Jornada (string equipo_local, string equipo_visitante, int jornada, long fecha_evento)	

	Descripción	Servicio web para la obtención de la emoción global extraída para cada equipo (local y visitante) en el contexto de jornada junto a los <i>tweets</i> asociados a dicha emoción .
	Precondiciones	<ul style="list-style-type: none"> Los argumentos deben cumplir lo siguientes requisitos: <ul style="list-style-type: none"> Equipo local y visitante no pueden ser nulos El valor de jornada y fecha_evento debe ser mayor a 0. Se debe haber obtenido previamente la información relativa al evento deportivo.
	Postcondiciones	<p>La función devuelve un objeto con los siguientes campos:</p> <ul style="list-style-type: none"> Código del estado de la petición Listado de objetos que recoge la información sobre la detección de emociones para cada equipo (Emoción extraída y Listado de <i>tweets</i> asociados)
	get_Emocion_Resultado (string equipo_local, string equipo_visitante, int jornada, long fecha_evento, int resultado)	
	Descripción	Servicio web para la obtención de la emoción global extraída para cada equipo (local y visitante) teniendo en cuenta el contexto objetivo, es decir el resultado, junto a los <i>tweets</i> asociados a dicha emoción .
	Precondiciones	<ul style="list-style-type: none"> Los argumentos deben cumplir lo siguientes requisitos: <ul style="list-style-type: none"> Equipo local y visitante no pueden ser nulos El valor de jornada y fecha_evento debe ser mayor a 0. El valor de resultado debe ser 1 (gana equipo local), 0 (empate) o -1 (gana equipo visitante) Se debe haber obtenido previamente la información relativa al evento deportivo.

	Postcondiciones	<p>La función devuelve un objeto con los siguientes campos:</p> <ul style="list-style-type: none"> • Código del estado de la petición • Listado de objetos que recoge la información sobre la detección de emociones para cada equipo (Emoción extraída y Listado de <i>tweets</i> asociados)
	get_Porcentaje_Emociones (string equipo, int jornada, int métrica)	
	Descripción	Servicio web para la obtención del porcentaje de recursos asociados a cada emoción estudiada para los parámetros equipo, jornada y tipo de análisis (métrica).
	Precondiciones	<ul style="list-style-type: none"> • Los argumentos deben cumplir lo siguientes requisitos: <ul style="list-style-type: none"> ○ Equipo no puede ser nulo ○ El valor de jornada debe ser mayor a 0. ○ El valor de métrica debe ser 1 (contexto jornada), 2 (contexto semanal) o 3 (considerando contexto objetivo (resultado)) • Se debe haber obtenido previamente la información relativa al evento deportivo. • Se debe haber realizado previamente el análisis de emociones para el equipo, jornada y métrica definidos y existir una observación de base de datos asociada.
	Postcondiciones	<p>La función devuelve un objeto con los siguientes campos:</p> <ul style="list-style-type: none"> • Código del estado de la petición • Objeto que contienen un listado de porcentajes asociados a las emociones extraídas.

Tabla 40. Componente *EmotionController*

Componente	NewsExtractor	
Propósito	Componente encargado de la obtención, a través del servicio externo proveedor de información deportiva, de la información relativa a los eventos deportivos objeto y del acceso a la base de datos ArangoDB para el almacenamiento y consulta de dicha información.	
Dependencias		
Interfaz	iNewsExtractor	
Contrato de operaciones	Extraer_informacion_Evento (List <string> equipos, int jornada)	
	Descripción	Método cliente encargado de consumir el servicio web del proveedor de información deportiva para la obtención de la información de los eventos asociados a los equipos y jornada especificados.
	Precondiciones	<ul style="list-style-type: none"> Los argumentos deben cumplir lo siguientes requisitos: <ul style="list-style-type: none"> El listado de equipo no puede ser nulo El valor de jornada debe ser un número de jornada válido. Debe existir un <i>feed</i> del proveedor externo de información deportiva disponible para su consumo y que este contenga la información consultada.
	Postcondiciones	<p>La función devuelve un listado de objetos con la información relativa a los eventos asociados a los equipos y jornada especificados. Los campos de dicho objeto son los siguientes:</p> <ul style="list-style-type: none"> Equipo local Equipo visitante Jornada

		<ul style="list-style-type: none"> Fecha y hora del partido Resultado
	guardar_Evento (Objeto evento)	
	Descripción	Método encargado del almacenamiento en base de datos del evento especificado.
	Precondiciones	<p>Se debe haber obtenido el objeto evento previamente mediante el servicio proveedor de información deportiva y que este contenga los siguientes campos:</p> <ul style="list-style-type: none"> Equipo local Equipo visitante Jornada Fecha y hora del partido Resultado
	Postcondiciones	El objeto Evento se encuentra almacenado en base de datos.
	Get_Evento (string equipo, int jornada)	
	Descripción	Método encargado de la consulta en base de datos del evento asociado al equipo y jornada especificados.
	Precondiciones	Deben existir eventos registrados en base de datos.
	Postcondiciones	Se obtiene el objeto evento consultado.
	Get_Ultima_Jornada ()	
	Descripción	Método de consulta en base de datos sobre el número de la última jornada sobre la que se tiene eventos deportivos almacenados.

	Precondiciones	Deben existir datos registrados en base de datos sobre los eventos deportivos objeto de estudio.
	Postcondiciones	Se obtiene un dato de tipo <i>int</i> que indica la última jornada almacenada.

Tabla 41. Componente NewsExtractor

Componente	EmotionExtractor	
Propósito	Componente encargado de las tareas análisis de emociones de los recursos asociados a los equipos y jornada objeto de estudio	
Dependencias	iCake	
Interfaz	iEmotionExtractor	
Contrato de operaciones	Extraer_Emoción_Semanal (string equipo, long fecha_evento, int jornada)	
	Descripción	Método encargado de la detección de emociones en los recursos asociados al equipo especificado y que hayan sido publicados desde el inicio de jornada hasta la fecha de inicio del evento (contexto semanal) y del almacenamiento de dicho resultado.
	Precondiciones	No debe existir en base de datos observaciones de resultado de análisis relativas al contexto semanal para el equipo y jornada especificados.
	Postcondiciones	<ul style="list-style-type: none"> La función devuelve un objeto con los siguientes campos: <ul style="list-style-type: none"> Emoción global extraída para el contexto jornada Listado de <i>tweets</i> asociados Se encuentra almacenado en base de datos el resultado del análisis para el contexto semanal
	Extraer_Emoción_Jornada (string equipo, long fecha_evento, int jornada)	
	Descripción	Método encargado de la detección de emociones en los recursos asociados al equipo especificado y que hayan sido publicados durante el tiempo en que transcurre el evento (contexto jornada) y del almacenamiento de dicho resultado.

	Precondiciones	No debe existir en base de datos observaciones de resultado de análisis relativas al contexto jornada para el equipo y jornada especificados.
	Postcondiciones	<ul style="list-style-type: none"> La función devuelve un objeto con los siguientes campos: <ul style="list-style-type: none"> Emoción global extraída para el contexto jornada Listado de <i>tweets</i> asociados Se encuentra almacenado en base de datos el resultado del análisis para el contexto jornada
	Extraer_Emoción_ContextoObjetivo (string equipo, long fecha_evento, int jornada, int resultado)	
	Descripción	Método encargado de la detección de emociones en los recursos asociados al equipo especificado y que hayan sido publicados desde el inicio de jornada hasta la fecha de fin del evento, utilizando el resultado del evento (contexto objetivo) en el algoritmo de clasificación, y, además, lleva a cabo el almacenamiento de dicho resultado.
	Precondiciones	No debe existir en base de datos observaciones de resultado de análisis relativas al contexto objetivo para el equipo y jornada especificados.
	Postcondiciones	<ul style="list-style-type: none"> La función devuelve un objeto con los siguientes campos: <ul style="list-style-type: none"> Emoción global extraída para el contexto jornada Listado de <i>tweets</i> asociados Se encuentra almacenado en base de datos el resultado del análisis para el contexto objetivo.
	Get_Recurso (string equipo, long fecha_ini, long fecha_fin)	
	Descripción	Método de consulta de base de datos de los recursos asociados al equipo especificado y que hayan sido publicados entre las fechas fecha_ini y fecha_fin.

	Precondiciones	Deben existir datos registrados en base sobre los recursos consultados.
	Postcondiciones	Se devuelve un listado de recursos asociados al equipo y fechas especificadas
	Get_Observaciones_Emocion (string equipo, int jornada, string métrica)	
	Descripción	Método de consulta de la observación resultado de la emoción global detectada para el equipo, jornada y estrategia de análisis (métrica) especificadas.
	Precondiciones	<ul style="list-style-type: none"> Los argumentos deben cumplir lo siguientes requisitos: <ul style="list-style-type: none"> Equipo no puede ser nulo El valor de jornada debe ser mayor a 0. El valor de métrica debe ser 1 (contexto jornada), 2 (contexto semanal) o 3 (considerando contexto objetivo (resultado)) Se debe haber realizado previamente el análisis de emociones para el equipo, jornada y métrica definidos y existir una observación de base de datos asociada.
	Postcondiciones	La función devuelve un objeto con los siguientes campos: <ul style="list-style-type: none"> Emoción global extraída. Listado de <i>tweets</i> asociados
	Get_Observaciones_PorcentajesEmocion (string equipo, int jornada, int métrica)	
	Descripción	Método de consulta de la observación sobre el porcentaje de recursos asociado a cada emoción estudiada para los parámetros equipo, jornada y estrategia de análisis (métrica).
	Precondiciones	<ul style="list-style-type: none"> Los argumentos deben cumplir lo siguientes requisitos:

		<ul style="list-style-type: none"> ○ Equipo no puede ser nulo ○ El valor de jornada debe ser mayor a 0. ○ El valor de métrica debe ser 1 (contexto jornada), 2 (contexto semanal) o 3 (considerando contexto objetivo (resultado)) • Se debe haber realizado previamente el análisis de emociones para el equipo, jornada y métrica definidos y existir una observación de base de datos asociada.
	Postcondiciones	La función devuelve un listado con el porcentaje de recursos asociados a cada emoción resultado del análisis
	Extraer_Emoción_Recurso (Objeto recurso)	
	Descripción	Método encargado del análisis de emoción a nivel de recurso.
	Precondiciones	El recurso de entrada no puede ser nulo.
	Postcondiciones	Devuelve la emoción mayoritaria presente en el recurso.

Tabla 42. Componente EmotionExtractor

Componente	Cake	
Propósito	Componente que implementa la API de la herramienta Knowledge Manager para el acceso a las funcionalidades de consulta y gestión de la ontología utilizadas para el análisis de emociones y las tareas de indexación y encaje de patrones.	
Dependencias		
Interfaz	iCake	
Contrato de operaciones	Indexar_texto(string texto_recurso)	
	Descripción	Método de indexación del texto de entrada que permite la aplicación de la técnicas de <i>pattern matching</i> para la detección de emociones en el texto.
	Precondiciones	El texto de entrada no puede ser nulo.
	Postcondiciones	La función devuelve el listado de patrones o <i>keywords</i> asociados a emoción definidos en la ontología que encajan con el texto indexado.

Tabla 43. *Componente Cake*

4.3. Proceso de detección de emociones en los tweets

Como se ha comentado anteriormente, el sistema de análisis diseñado cuenta con un enfoque basado en conocimiento en el que se puede distinguir dos partes: una base de conocimiento para el reconocimiento de emociones, que está conformada por el léxico de emociones SEL, un conjunto de *emojis* etiquetados con emociones definido en “Emoji as Emotion Tags for Tweets” [22] y un conjunto de patrones, y, por otro lado, un motor de inferencia basado en reglas. A continuación, se muestra el diseño del proceso de detección de emociones en los *tweets*:

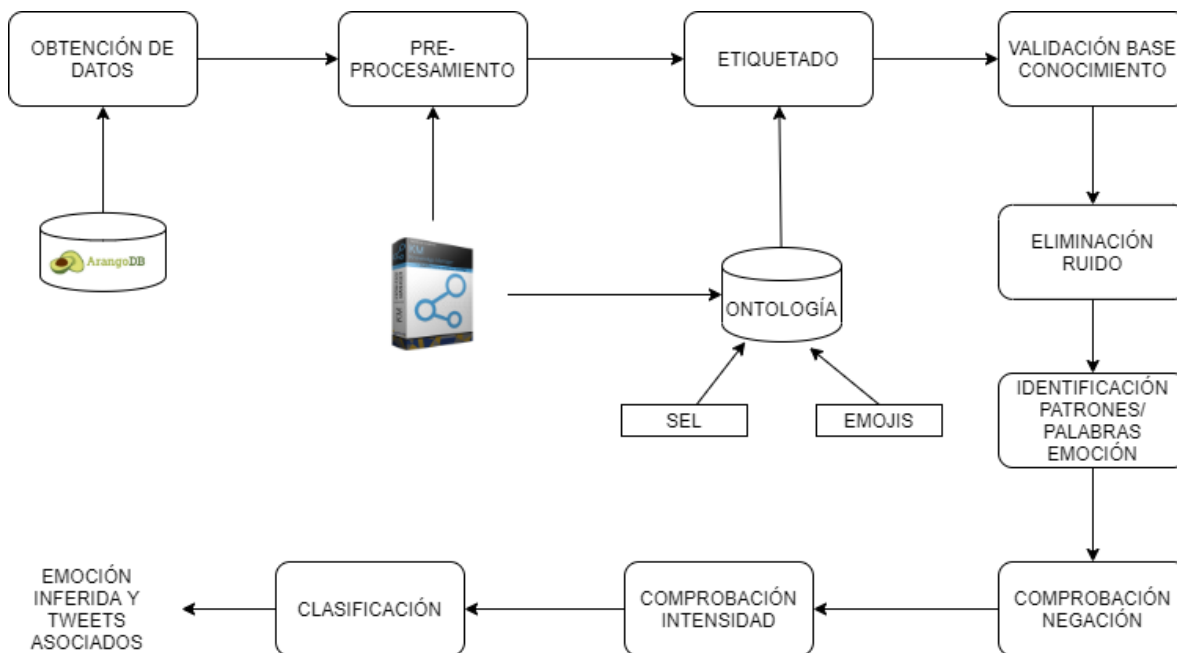


Figura 8. Esquema general de la arquitectura funcional del sistema de análisis

La descripción de los bloques funcionales del proceso de análisis observados en la Figura 8 es la siguiente:

- **Obtención de datos:** proceso encargado de consultar de base de datos ArangoDB los recursos de Twitter extraídos por la plataforma SMART-RTVE para su posterior análisis. En la consulta de dichos recursos se filtran en base al equipo y jornada a estudiar y se obtienen el texto de los recursos filtrados (*tweet*).
- **Preprocesamiento:** bloque que realiza la transformación del texto de entrada a un formato adecuado para su posterior indexación. Los procesos incluidos en la transformación son la tokenización, normalización y etiquetado sintáctico del texto. Este bloque utiliza las funciones ofrecidas por Knowledge Manager para llevar a cabo dichos procesos de tratamiento del texto.
- **Etiquetado:** proceso que se encarga de indexar los textos procesados de acuerdo al conocimiento definido (ontología y patrones) en la herramienta Knowledge Manager para la detección de emociones en estos.
- **Validación base conocimiento:** bloque que comprueba la existencia de patrones válidos en el resultado de la indexación.

- **Eliminación ruido:** proceso que identifica los patrones que indican que el texto hace referencia a información no relevante para el análisis (información sobre otros eventos u otras categorías de equipo) para su posterior eliminación del proceso.
- **Identificación de palabras/patrones emocionales:** parte del sistema encargado de identificar aquellos patrones encontrados en el texto que están asociados a una emoción en concreto y contabilizar su ocurrencia.
- **Comprobación negación:** proceso de identificación de aquellos patrones que detecten la presencia de una emoción de forma negada, para evitar la contabilización de esta misma.
- **Comprobación intensidad:** bloque encargado de identificar el nivel de fuerza de la emoción detectada, modificando el valor de ocurrencia en función de si las palabras emocionales aparecen acompañadas por intensificadores o mitigadores.
- **Clasificación:** proceso que se encarga de la inferencia de la emoción correspondiente a cada *tweet*, mediante la contabilización del número de ocurrencias de cada emoción y el peso de las mismas, siendo la emoción con mayor peso asociado la emoción inferida para dicho recurso. Una vez extraída la emoción de cada *tweet*, se determina el porcentaje de *tweets* asociados a cada emoción y se establece como emoción global aquella emoción con mayor porcentaje. Para el caso de la estrategia de análisis teniendo en cuenta el resultado (contexto objetivo) se modifica dicho porcentaje en función del resultado, potenciando las emociones positivas si este es favorable o, en caso de resultado desfavorable, las emociones negativas.

4.4. Ontología que conforma la base de conocimiento

A continuación se especifica la definición y configuración de la ontología a emplear para modelar la información relativa a la extracción de emociones que será utilizada como base de conocimiento para el problema de clasificación.

Para dicha definición, previamente se ha estudiado los recursos existentes para la detección de emociones en corpus en español en redes sociales. Tras dicho estudio, donde destaca la

escasez de recursos en castellano, se han escogido dos tipos de recursos: léxico de emociones y *emojis* etiquetados con emoción.

El léxico de emociones utilizado es el “Spanish Emotional Lexicon (SEL)” [26] que se trata de un léxico creado por Grigori Sidorov en el Instituto Politécnico Nacional de México. Este léxico está conformado por 2.036 palabras en español que tienen asociada al menos una emoción con un factor de probabilidad de afectividad (PFA), siendo las categorías de emoción utilizadas las emociones básicas de Ekman: alegría, ira, miedo, tristeza, sorpresa y aversión. El PFA de cada palabra del léxico viene determinado por el etiquetado por parte de 19 anotadores del nivel de certeza respecto a cada emoción (nulo, bajo, medio y alto). En la siguiente figura se muestra un ejemplo de un conjunto de palabras que conforma el SEL y las probabilidades asociadas.

Palabra	Nula[%]	Baja[%]	Media[%]	Alta[%]	PFA	Categoría
abundancia	0	0	50	50	0.83	Alegría
acabalar	40	0	60	0	0.396	Alegría
acatar	50	40	10	0	0.198	Alegría
acción	30	30	30	10	0.397	Alegría
aceptable	0	20	80	0	0.596	Alegría

Figura 9. Ejemplo de Spanish Emotional Lexicon (SEL) [26]

En este trabajo, tras la revisión de dicho vocabulario, se han filtrado aquellos términos con mayor PFA asociado a una emoción ($PFA > 0,5$), se ha realizado el etiquetado sintáctico de las mismas y se han adaptado al español propio del territorio de España aquellas palabras encontradas que pertenecen al dialecto de América Latina. De esta forma se han definido parte de los términos correspondientes a las emociones estudiadas que conformarán la ontología.

Además de los términos del léxico anteriormente mencionado, también se incluirán en la ontología un conjunto de *emojis* asociados a emociones. Este tipo de recurso resulta de gran importancia ya que los *emojis* son altamente utilizados por los usuarios de redes sociales para expresar su estado emocional. Como fuente de este tipo de recurso se ha utilizado el trabajo

“Emoji as Emotion Tag For Tweets” [22] que consiste en un conjunto de *emojis* anotados con la emoción correspondiente, como en el caso del léxico anterior, basándose en el modelo de emociones de las 6 categorías básicas de Ekman, tal como aparecen en la siguiente figura.



Figura 10. Clasificación de emojis por emoción [22].

Posteriormente, para el refinamiento de los resultados, se ha decidido incluir ciertos términos basados en jerga del dominio deportivo, en concreto del ámbito futbolístico. La selección de dichos términos se ha basado en la observación de aquellos *tweets* pertenecientes al ámbito objeto de estudio. Como ejemplo de este tipo de vocabulario se pueden destacar palabras como ‘partidazo’ o ‘golazo’ que han sido asociados a la emoción alegría.

En la Tabla 44 se muestra el número de términos que conforman cada categoría de emoción.

Emoción	N.º Términos	Porcentaje
Alegría	385	32,93%
Tristeza	205	17,54%
Sorpresa	88	7,53%
Ira	217	18,56%
Aversión	147	12,58%
Miedo	127	10,86%

Tabla 44. Distribución de términos por emoción

En base al vocabulario definido anteriormente y su agrupación en clústeres en función de la emoción asociada, se han diseñado una serie de patrones para la detección de los siguientes aspectos: *keywords* de emociones (basados en el vocabulario anterior), *keywords* para la detección de ruido, estructuras de negación, intensificadores y mitigadores.

A continuación, se presenta los patrones creados para la detección de estructuras de negación, intensificación y mitigación, respectivamente, mediante la especificación de los *slots* que conforman la estructura del patrón y la enumeración de una serie de ejemplos asociados a las emociones a identificar. Cabe apuntar que la aparición de *slots* entre símbolos de interrogación ‘¿<Slot>?’ indica que su presencia en la estructura es opcional.

Patrones negación

Patrón negación emoción 1	
Estructura	<Adverbio Negación o Preposición Sin> ¿<Pronombre>? ¿<Pronombre>? ¿<Verbo Haber>? <Verbo Emoción>
Ejemplos	<ul style="list-style-type: none"> Ejemplo Alegría: No le ha alegrado. Ejemplo Sorpresa: Sin sorprender.
Patrón negación emoción 2	
Estructura	<Pronombre Negación> ¿<Pronombre>? ¿<Pronombre>? ¿<Verbo Haber>? <Verbo Emoción>
Ejemplos	<ul style="list-style-type: none"> Ejemplo Tristeza: Ninguno ha decepcionado. Ejemplo Miedo: Nadie lo teme.
Patrón negación emoción 3	
Estructura	<Determinante Negación> <Adjetivo Emoción> <Nombre >.
Ejemplos	<ul style="list-style-type: none"> Ejemplo Ira: Ningún hostil gesto. Ejemplo Aversión: Ningún asqueroso momento.
Patrón negación emoción 4	
Estructura	<Determinante Negación> <Nombre > <Adjetivo Emoción>
Ejemplos	<ul style="list-style-type: none"> Ejemplo Ira: Ningún gesto enfadado. Ejemplo Miedo: Ninguna jugadora atemorizada.
Patrón negación emoción 5	
Estructura	<Determinante Negación o Preposición Sin > <Nombre Emoción>
Ejemplos	<ul style="list-style-type: none"> Ejemplo Alegría: Sin Éxito. Ejemplo Tristeza: Ninguna Derrota.
Patrón negación emoción 6	
Estructura	<Adverbio Negación o Pronombre Negación> ¿<Verbo Haber-Tener>? <Verbo ser o estar> <Adjetivo Emoción>
Ejemplos	<ul style="list-style-type: none"> Ejemplo Sorpresa: No ha estado apoteósico. Ejemplo Aversión: Nadie ha sido desagradable.

Tabla 45. Patrones para la detección de negación.

Patrones intensificador

Patrón intensificador emoción 1	
Estructura	<Determinante Intensificador o Adjetivo Comparador Intensificador > <Nombre Emoción>
Ejemplos	<ul style="list-style-type: none"> • Ejemplo Alegría: Mucha alegría. • Ejemplo Tristeza: Más tristeza.
Patrón intensificador emoción 2	
Estructura	<Adverbio Intensificador > <Adjetivo Emoción>
Ejemplos	<ul style="list-style-type: none"> • Ejemplo Ira: Tan detestable. • Ejemplo Aversión: Bastante horroroso.
Patrón intensificador emoción 3	
Estructura	<Adverbio Intensificador > <Adverbio Emoción>
Ejemplos	<ul style="list-style-type: none"> • Ejemplo Sorpresa: Muy sorprendentemente. • Ejemplo Miedo: Tan despiadadamente.
Patrón intensificador emoción 4	
Estructura	<Verbo Emoción > <Adverbio Intensificador >
Ejemplos	<ul style="list-style-type: none"> • Ejemplo Alegría: Divierte mucho. • Ejemplo Tristeza: Ha decepcionado bastante.

Tabla 46. Patrones para la detección de intensificadores.

Patrones mitigador

Patrón mitigador emoción 1	
Estructura	<Determinante Mitigador o Pronombre Mitigador o Adjetivo Comparador Mitigador> <Nombre Emoción>
Ejemplos	<ul style="list-style-type: none"> • Ejemplo Alegría: Poca alegría. • Ejemplo Miedo: Algo de alarma.
Patrón mitigador emoción 2	
Estructura	<Adverbio Mitigador > <Adjetivo Emoción>
Ejemplos	<ul style="list-style-type: none"> • Ejemplo Ira: Menos detestable. • Ejemplo Sorpresa: Algo asombroso.
Patrón mitigador emoción 3	
Estructura	<Adverbio Mitigador > <Adverbio Emoción>
Ejemplos	<ul style="list-style-type: none"> • Ejemplo Alegría: Menos alegremente. • Ejemplo Aversión: Menos repulsivamente.
Patrón mitigador emoción 4	
Estructura	<Verbo Emoción > <Adverbio Mitigador >
Ejemplos	<ul style="list-style-type: none"> • Ejemplo Tristeza: Ha decepcionado algo. • Ejemplo Sorpresa: Ha maravillado poco.

Tabla 47. Patrones para la detección de mitigadores.

5. Implementación

Una vez presentado el diseño del sistema, en esta sección se explica la implementación del sistema de análisis emociones, detallando como se ha desarrollado cada módulo que lo conforma.

Las partes a distinguir en la explicación del proceso de desarrollo son las siguientes:

- Módulo de obtención de recursos
- Módulo de construcción de la ontología: configuración del componente Knowledge Manager.
- Módulo de obtención de contexto objetivo
- Módulo de análisis y clasificación de emociones
- Submódulo de Social Media News de muestra de resultados.

5.1. Módulo de obtención de recursos

En este módulo se ha desarrollado la lógica de obtención de los recursos (*tweets*) para su posterior análisis para la detección de la emoción asociada. Este proceso se realiza mediante la plataforma SMART-RTVE que almacena en la base de datos ArangoDB los recursos de las cuentas de Twitter monitorizadas y permite su posterior visualización en la misma plataforma.

En primer lugar, se ha realizado una configuración manual en SMART-RTVE de las cuentas a monitorizar en Twitter. Para ello, ha sido necesario un estudio previo sobre qué cuentas y *hashtags* son de interés en relación con los equipos objeto de análisis. Esta tarea resulta de gran importancia ya que en ella reside la confianza del contexto subjetivo extraído, por lo que ha sido necesario identificar qué cuentas ofrecen una información representativa del público objetivo.

Para la elección de cuentas y *hashtags* se ha tenido en cuenta cuáles de ellos publican o están asociados a contenido relacionado con los equipos de interés. En el caso de las cuentas se han seleccionado cuentas oficiales, de seguidores y de medios locales y nacionales. En la siguiente tabla se presentan las cuentas y *hashtags* monitorizados.

	Cuentas					Hashtags			
	Oficial	Seguidores	Medios de comunicación (locales)	Medios de comunicación (comunes)	Cuentas deportivas (comunes)	Oficiales	Seguidores	Partido	Liga
CF Talavera de la Reina	@CFTalavera_	@TalaveraStats @taladefender	@tribunadetoledo @eldigitalCLM @lavozdetalavera			#OrgulloCerámico #VamosTala			
Villarrobledo	@CPVillarrobledo	@ultratroyanos @CurvaLobo @cpvinfotbol	@SERDeportivosLM			#MuchoVilla #JuntosSomosMasFuer tes			
Villarrubia	@formacvillarrub	@FrenteVillarru2 @Jesus_Guti	@LaTribunaCR @eldigitalCLM @Lanzadigital			#IlusionBlanquiazul #LaIlusionDeUnPueblo			
FC Cartagena	@FCCartagena_efs	@pantaniefese @fpfct @EfesistaOficial	@laverdad_ct @diariolaopinion			#vamosCartagena #FCCartagena	#Efesé		
San Fernando CD	@SanFernando_CD	@PalanteCDSF @SFCODYMAS	@RadioCadiz @DeportesCopeCAD			#UniónYFuerza #RecreSanFernando	#VamosSanfer		
Marbella	@marbella_fc	@Botijomarbella @MarbellerosMFC	@CopeMarbella @RTVMarbella			#MarbellaFC #NuevoMarbella	#VamosMiMarbella		
CD Badajoz	@CDBadajoz	@2aDivisionB_CDB @CDBadajozStats @EternoCDB	@hoyextremadura			#AupaCDB #unahistoriaquecontar	#AupaBada #DaleBadajoz		
Cádiz B	@Cadiz_CFCantera	@diariocadiz				#CádizB			
Yeclano Deportivo	@Yeclano_dptvo	@PCurvabar	@ORMurcia @periodicoyecla @YeclaSport			#SomosYeclano	#yeclano		
Córdoba	@Cordoba_CF_ofi	@Cordobacsad @pcpcruzconde	@PTVDeportesCord @cordopolis_es @CordobesismoCom	@SegundaUniv erso	@2Bgoles @SegundaB_G3 @SegundaB_G4 @Jugadores2ab @RFAF	#VolveremosCCF #CórdobaCF			
R. Balompédica Linense	@RBL1912	@BalonaDatos @DeLaBalona @SoloBalona	@europa_sur @DIARIO_AREA	@diarios @marca @footers		#QuieroBalona #SomosLaBalona	#SomosLaRecia #vamosbalona	#'EquipoLocal'+ 'EquipoVisitante'	#2bg4 #grupoIV #SegundaB #2BGIV
Sevilla Atlético	@CanteraSFC	@Orgullo_Nervion	@LaCanteraSFCRad			#VamosSAT #CanteraSFC	#SevillaAtlético.		
Don Benito	@Cddonbenito1928	@CristianDiazPa9 @pitysanchez @FrenteOpuesto	@hoyextremadura			#DaleDepor #CDDonBenito			
Recreativo	@recreoficial	@jozaprax @piyaorecre @La_decanoteca @Sos_Recre	@HuelvaBN @AytoHuelva @vivahuelva			#TellevoenelAlma	#YoSoydelRecre		
UCAM Murcia	@UCAMMurciaCF		@InterDeportivo @AchoPasala @diariolaopinion			#UCAMMurciaCF			
Real Murcia	@realmurciacsad	@jorgefenor @FacultadGrana @RegionGrana @GA_Murcianista	@murciaplaza_com			#RMUseSiente #RealMurcia	#RugeRealMurcia		
Atlético Sanluqueño	@atcosanluqueno	@Marea_Verde_05	@csurDeportes			#ASCF1948 #OrgulloSanluqueño	#Sanluqueño		
Recreativo de Granada	@CanteraNazari	@Grada1931GCF	@COPECordoba @granadaenjuego			#CanteraNazari #RecreativoGranada #UnRomanoNuncaSeR inde	#RecreativoGranad aCCF #MéridaAD		
Mérida	@Merida_AD	@10Tro10 @AnimaRomano	@el_fdc @DeportesCexTV						
Algeciras	@AlgecirasCF		@minutoes @europa_sur			#AlgecirasCF			

Figura 11. Cuentas y Hashtags de Twitter monitorizados.

Tras la configuración manual de las cuentas y *hashtags* a monitorizar, a través de la plataforma SMART-RTVE, se pueden visualizar diferentes aspectos de la actividad de estas, como son el número de publicaciones, seguidores y seguidos, tipología de recursos (*tweets*, imágenes o vídeos) y estadísticas sobre el número de recursos publicados por hora y la variación de número de seguidores. La información visualizada en la plataforma SMART-RTVE se muestra en la siguiente figura.

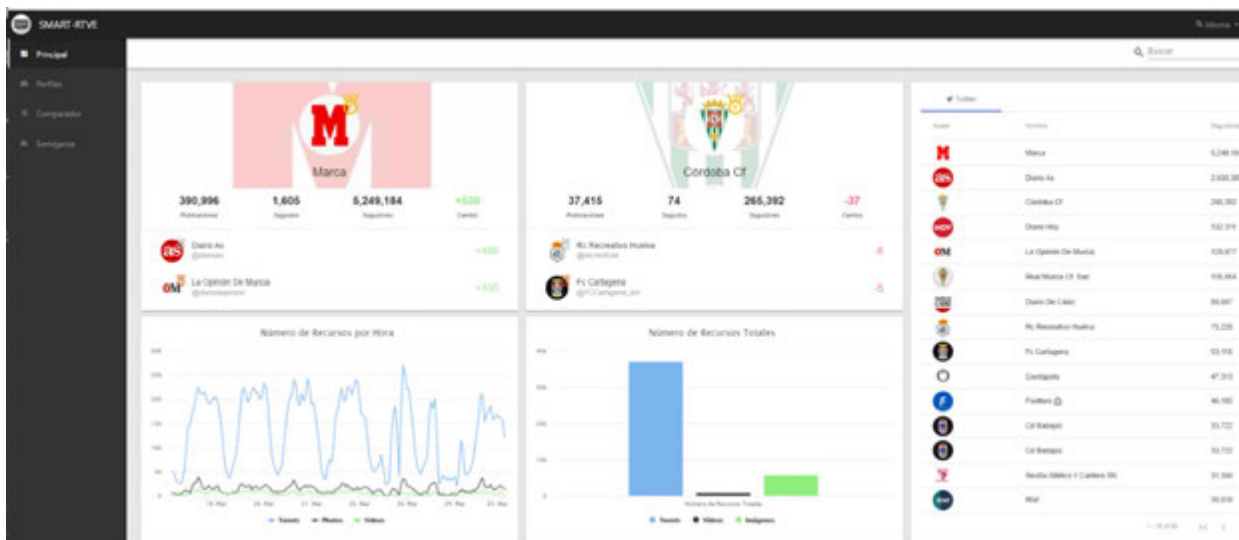


Figura 12. Pantalla de monitorización de cuentas de la plataforma SMART-RTVE.

Una vez establecida dicha configuración manual, se realiza la implementación del proceso de obtención de recursos en el sistema. Este proceso, en primer lugar, se conecta con la base de datos ArangoDB donde son almacenados los recursos recogidos por la plataforma SMART-RTVE. En la Figura 13 y Figura 14 se muestran, respectivamente, las colecciones de base de datos donde se encuentran los recursos a analizar (colección ‘resources’) y la información de los usuarios generadores de los mismos (colecciones ‘rtve’ y ‘anonymous’) y, por otro lado, un ejemplo de recurso almacenado en formato JSON.

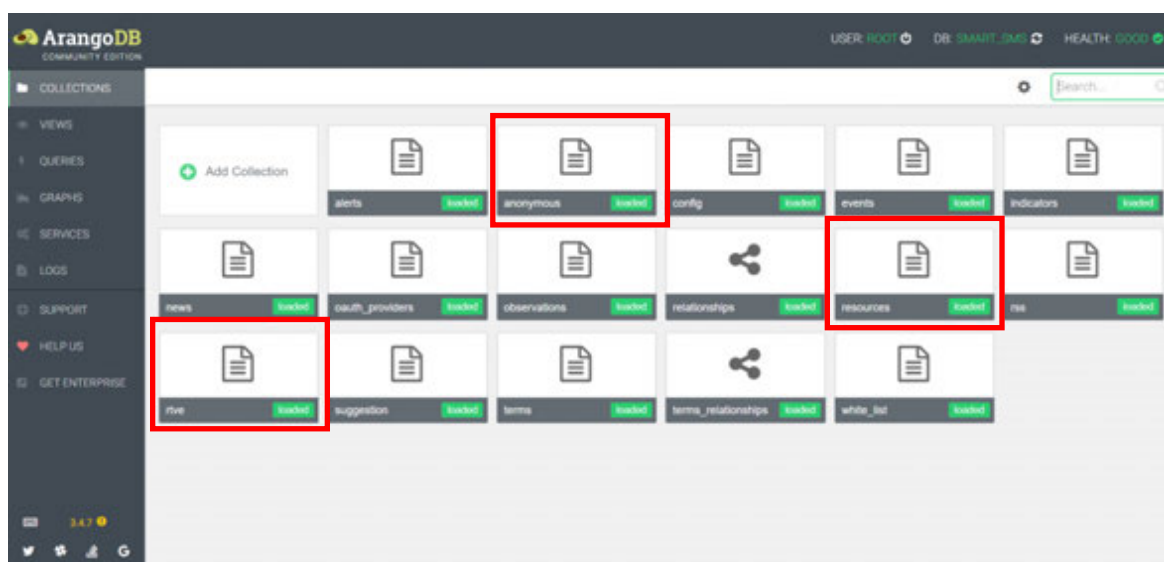


Figura 13. Colecciones para el almacenamiento de recursos e información de usuario de ArangoDB

```

{
  "_key": "5735204325247026999",
  "_id": "resources/5735204325247026999",
  "_rev": "_ZdfIAAq---",
  "source_id": "1186556557586075648",
  "type": "TWEET",
  "autor": "anonymous/1853071497637295451",
  "socialNetwork": "TWITTER",
  "createdDate": 1571739303000,
  "discoveredDate": 1571836477476,
  "numberLikes": 5,
  "numberMentions": 2,
  "numberResponses": 0,
  "mentions": [
    "rtve/4956208993060940016",
    "anonymous/-5762829623985323083"
  ],
  "responses": null,
  "mentionedBy": null,
  "respondedBy": null,
  "likedBy": null,
  "lenguaje": "es",
  "url": "https://twitter.com/LaTribunaCR/status/1186556557586075648",
  "description": "El usuario LaTribunaCR ha publicado este tweet. El usuario formacvillarrub ha retweeteado este tweet.",
  "text": "«En casa ya merecemos la victoria» @formacvillarrub @_JaviS5 https://t.co/DEUkPzBevw https://t.co/BU3kA4blDB",
  "keywords": null,
  "place": null,
  "placeNormalized": null,
  "action": "PUBLISH",
  "mediaURL": "http://pbs.twimg.com/media/EHd-5qtUUAhntf.jpg",
  "specialProperties": {
    "HASTAGS": [],
    "PHOTOS": [
      "http://pbs.twimg.com/media/EHd-5qtUUAhntf.jpg"
    ],
    "VIDEOS": [],
    "RETWEETED_BY": [
      "rtve/4956208993060940016"
    ]
  }
},

```

Figura 14. Ejemplo de recurso en formato JSON

Dado el objeto recurso presentado en la figura anterior, se muestran a continuación los campos utilizados en el desarrollo del sistema junto a su descripción:

- **source_id:** número de identificación de recurso.
- **createdDate:** fecha de publicación del recurso en Twitter
- **url:** URL de Twitter de acceso al *tweet*.
- **description:** descripción que indica la cuenta que ha publicado el *tweet* y las cuentas que lo han retweeteado
- **text:** texto del *tweet*.
- **specialProperties.HASTAGS:** listado de *hashtags* incluidos en el contenido del *tweet*.

Una vez establecida la conexión con la base de datos, se consultan mediante el lenguaje propio de ArangoDB (AQL) los recursos almacenados en función del equipo y jornada a analizar. Por tanto, se extraen aquellos *tweets* de las cuentas y *hashtags* asociados al equipo seleccionado (a través del campo “description” para el caso de recursos de cuenta y, en el caso de *hashtags*, a través del campo “specialProperties.HASTAGS”) y que tienen una fecha y hora de creación (“createdDate”) correspondiente a la jornada (desde la hora de inicio del partido hasta las 3 horas siguientes) para el contexto subjetivo de jornada y, para el contexto semanal, los recursos creados durante la semana de jornada (fecha de inicio de la semana en que transcurre la jornada hasta la hora de inicio del partido).

5.2. Módulo de construcción de la ontología: configuración del componente Knowledge Manager

Este módulo se encarga de la construcción de la ontología definida en el subapartado 4.4 Ontología que conforma la base de conocimiento para la detección de emociones mediante la configuración del componente software Knowledge Manager (KM). Antes de explicar cómo se ha realizado dicha configuración, cabe destacar que se utiliza como base una ontología en español proporcionada por *The Reuse Company* que se compone de más de 8.000 términos, todos ellos etiquetados sintácticamente. Esta ontología base será utilizada para la definición de patrones y a ella se añadirán los términos obtenidos para la detección de emociones.

La inclusión de los términos se realiza de forma manual, mediante la opción ofrecida por KM de importación desde Excel de términos, especificando la categoría sintáctica de cada uno de ellos.

Figura 15. Interfaz de Knowledge Manager para insertar lista de términos desde Excel.

Para que en la ontología quede recogida la emoción asociada a cada termino, se ha creado un clúster para cada emoción y se han incluidos los términos en el clúster correspondiente.

Término	Etiqueta sintáctica	Clúster	Tipo de relación	Idioma
Grandioso	ADJETIVO	«Sorpresa», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Estupendo	ADJETIVO	«Sorpresa», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Maravillosamente	ADVERBIO	«Sorpresa», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Admirablemente	ADVERBIO	«Sorpresa», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Admirable	ADJETIVO	«Sorpresa», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Fascinante	ADJETIVO	«Sorpresa», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Maravillar	VERBO	«Sorpresa», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Admirar	VERBO	«Sorpresa», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Festivo	ADJETIVO	«Emoción», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Contento	ADJETIVO	«Emoción», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Alegre	ADJETIVO	«Emoción», «Alegria», < Sin «Tipo de Relación Español (alfabetizació)»		
Impetu	NOMBRE	«Alegria», «u - NOMB» < Sin «Tipo de Relación Español (alfabetizació)»		
Benefactor	ADJETIVO	«Alegria», «tor - ADJET» < Sin «Tipo de Relación Español (alfabetizació)»		
Confort	NOMBRE	«Alegria», «t - NOMB» < Sin «Tipo de Relación Español (alfabetizació)»		

Figura 16. Interfaz de Knowledge Manager que muestra clúster <<Alegria>>

Para el caso de los términos propios de la jerga del dominio analizado, se ha decidido la creación del clúster 'Jerga' que contenga los clústeres hijos de nombre 'Jerga+Emoción' para

cada emoción estudiada. La decisión de incluir este tipo de términos mediante la creación de nuevos clústeres se debe a la necesidad de diferenciarlos por la especificidad del ámbito al que se refiere. De esta forma resulta de mayor facilidad su localización y su posible adaptación a otros ámbitos.

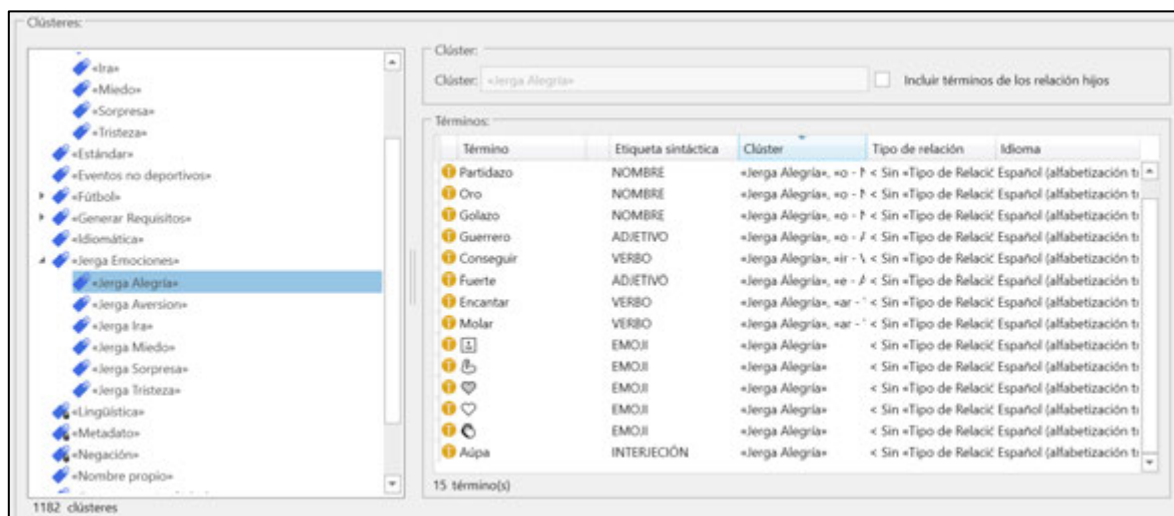


Figura 17. Interfaz de Knowledge Manager que muestra clúster <<Jerga Alegría>>

Además de los clústeres de emociones, en la ontología se han incluido clústeres con términos para la eliminación los recursos que suponen ruido para el análisis, es decir, aquellos *tweets* con información no relacionada con el evento y equipo objeto de estudio. La determinación de los términos que conforman estos clústeres ha sido en base al estudio manual de la información no relevante encontrada en los recursos recogidos. La inclusión de dichos términos también ha sido realizada mediante la interfaz de KM. Los clústeres creados han sido 2 y están conformados por términos relativos a eventos no relacionados y otras categorías de equipo.

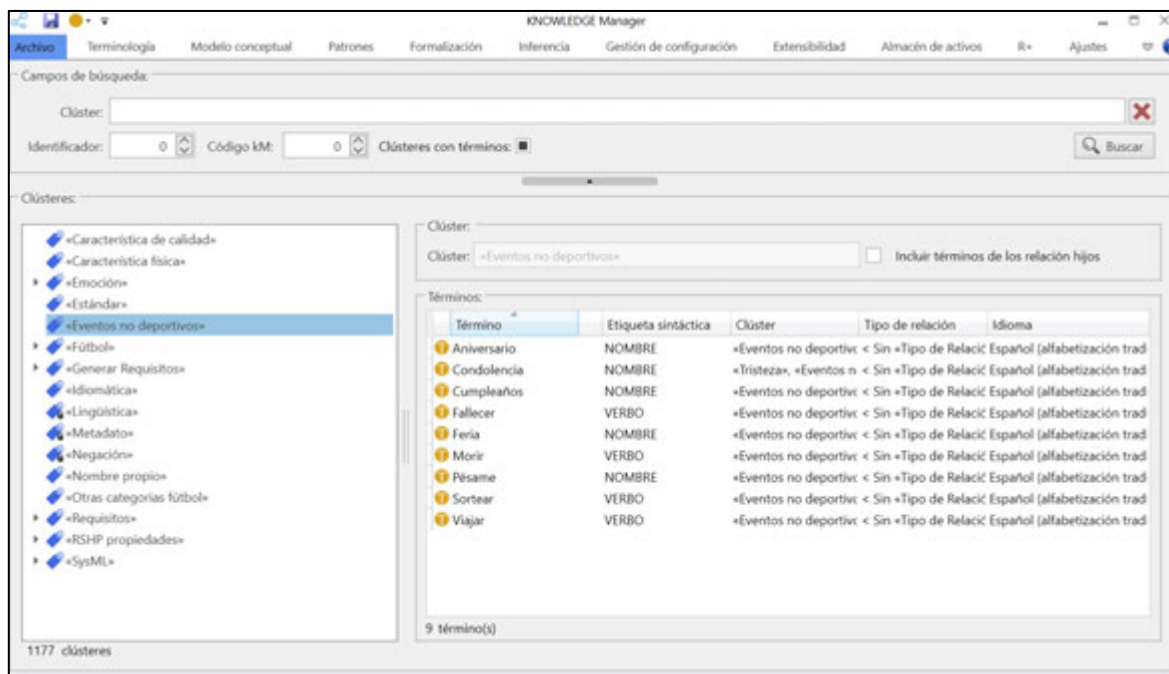


Figura 18. Interfaz de Knowledge Manager que muestra clúster <<Eventos no deportivos>>

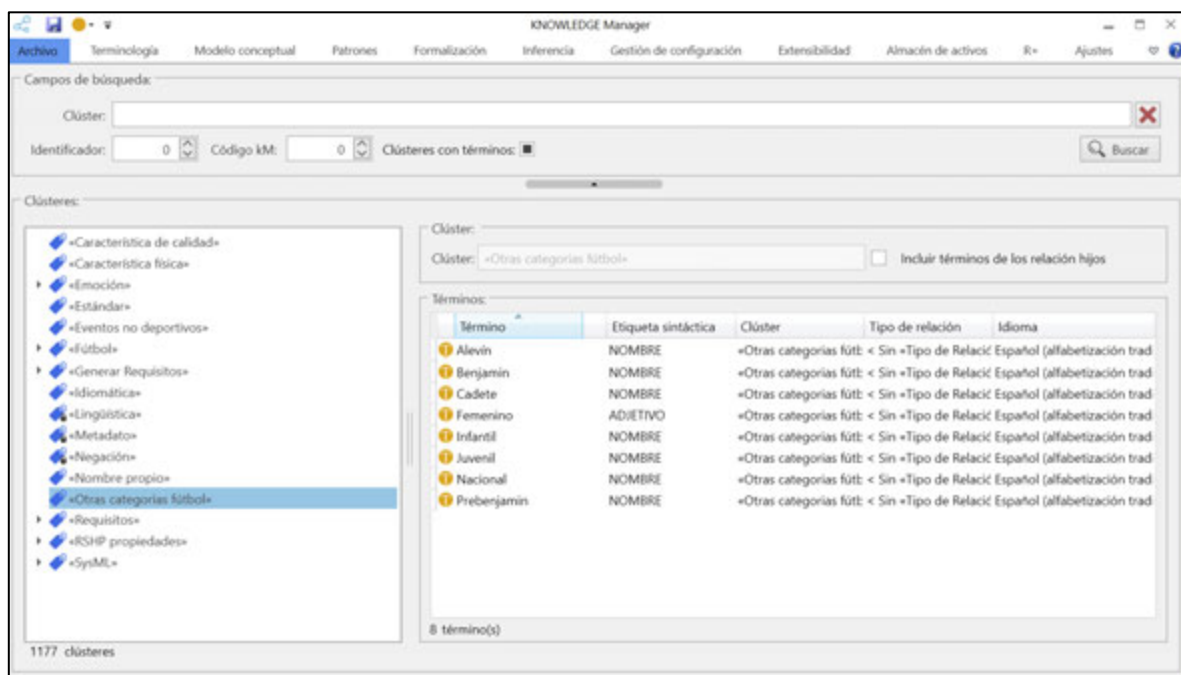


Figura 19. Interfaz de Knowledge Manager que muestra clúster <<Otras categorías fútbol>>

Por otro lado, la creación de patrones también se ha realizado, como para el caso del vocabulario, con la herramienta KM. Esta herramienta se encarga tanto de la gestión de términos como de patrones, lo que permite establecer una relación semántica entre los

patrones y los clústeres que incluyen el vocabulario creado, vinculando al menos uno de los elementos del patrón con los clústeres definidos.

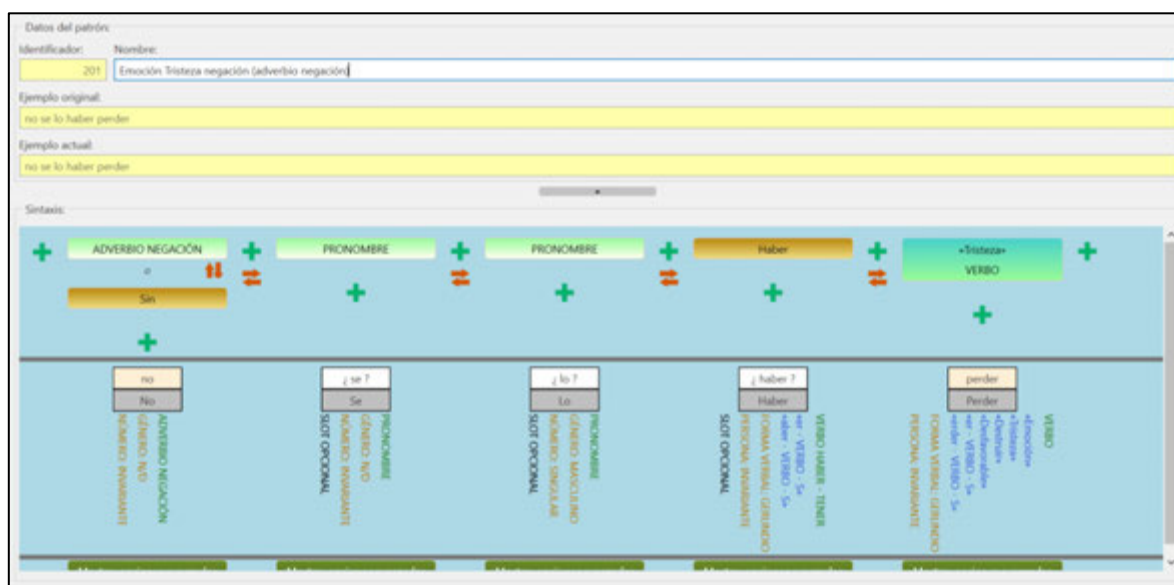


Figura 20. Interfaz de Knowledge Manager que muestra la configuración de patrones.

Los patrones creados, tanto los conformados por las *keywords* de emociones y las relativas a otros eventos y categorías, como por las estructuras de negación, intensificadores y mitigadores, serán consultados posteriormente en el módulo de análisis. Dicha consulta se realizará mediante el uso de técnicas de *pattern matching* en los *tweets* para la detección de emociones en estos recursos así como para la identificación de información no relevante para el estudio.

5.3. Módulo de obtención de contexto objetivo

Este módulo es el encargado de obtener la información referente a los partidos de fútbol de segunda división B grupo IV. Para ello, se utiliza un *feed* habilitado por el proveedor externo de información deportiva, el cual cuenta con una relación contractual con RTVE. Debido a que se trata de una API privada no es posible mostrar en el presente documento el tipo de respuesta obtenida del servicio.

En la implementación se ha creado un componente cliente en lenguaje C# que consume dicho servicio externo y procesa la información obtenida en formato JSON para su posterior almacenamiento en la colección 'news' en la base de datos ArangoDB.

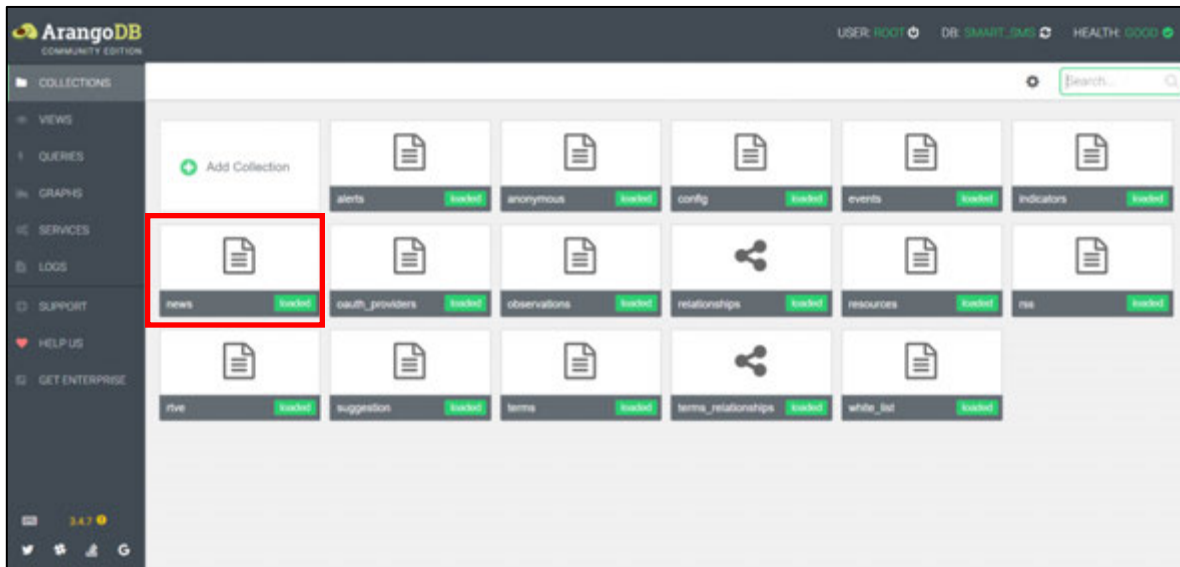


Figura 21. Colección para el almacenamiento de información deportiva en ArangoDB

La información obtenida mediante el servicio mencionado se utiliza para extraer el contexto objetivo de los partidos. Por tanto, dado un equipo y jornada seleccionados por el usuario, este módulo consulta en base de datos la información deportiva relativa al evento asociado a dicho equipo y jornada. El contexto objetivo extraído en este proceso consta de los siguientes datos:

- **match_id**: identificador del evento.
- **match_time**: fecha y hora del evento.
- **local_name**: nombre del equipo local.
- **visitor_name**: nombre del equipo visitante.
- **result**: resultado del partido en formato 'X-Y', donde X es el número de goles del equipo local e Y es el número de goles del visitante.
- **round**: número de jornada.

Por último, cabe destacar que se ha requerido la librería Newtonsoft.Json [50] en la implementación de este módulo para deserialización del objeto JSON obtenido como respuesta del servicio externo consultado.

5.4. Módulo análisis y clasificación de emociones

El módulo de análisis y clasificación de emociones, implementado en C#, es el responsable de la funcionalidad principal del sistema, llevando a cabo el proceso de inferencia de emociones que se encarga de la obtención de los recursos asociados a los parámetros especificados por el usuario (equipo y jornada) y su clasificación. El resultado de dicho análisis será proporcionado a la plataforma SMART-RTVE para su posterior visualización.

La extracción del contexto subjetivo de jornada y semanal, en función del equipo y jornada a analizar, se realiza a partir del módulo de obtención de recursos. Una vez obtenidos los recursos a analizar el sistema realizará una clasificación a 2 niveles. En primer lugar, mediante una metodología basada en conocimiento, se lleva a cabo el proceso de detección de emociones a nivel de recurso mediante el *matching* con los patrones basados en la ontología configurada que cuentan con las siguientes características: adjetivos, nombres, adverbios, verbos, *emojis* y estructuras de negación, intensificadores y mitigadores. En segundo lugar, una vez clasificados los recursos objeto de análisis, se realiza el proceso de inferencia de la emoción global.

Análisis a nivel de tweet

Tal como se especifica en el diseño del proceso de detección de emociones en los *tweets* (capítulo 4.3 Proceso de detección de emociones en los tweets), previamente a la aplicación de los algoritmos de clasificación de emociones, es necesario un preprocesamiento del texto y la aplicación de las técnicas de *pattern matching* en base a la ontología configurada. Para ello se utiliza la API Cake de la herramienta Knowledge Manager disponible en C#. El objeto resultado proporcionado es de tipo Dictionary<TKey,TValue>, que representa una colección de claves y valores. Dicho diccionario contienen el listado de patrones válidos y palabras claves encontradas en los recursos procesados.

Una vez obtenido dicho listado, se identificarán los patrones que indican la presencia de información no relevante en los recursos para la eliminación de estos ya que suponen ruido para el análisis. Los patrones que identifican este aspecto se encuentran compuestos por los clústeres mostrados en la Figura 18 y Figura 19.

Posteriormente, se identificarán aquellos patrones que detecten la presencia de emoción en el texto. Antes de contabilizar la ocurrencia de cierta emoción se comprueba si aparece de forma negada. Por ejemplo, el fragmento ‘No ha sido sorprendente’ contiene el adjetivo ‘sorprendente’ que se encuentra asociado al clúster de emoción <<*Sorpresa*>>, sin embargo el fragmento no expresa dicha emoción. Por tanto, es necesaria la identificación de dichos casos en el sistema para que no se contabilice la emoción negada como una ocurrencia de esta.

Tras la comprobación de negación, se aplica el algoritmo de clasificación para la detección de emociones y su intensidad. Dicho algoritmo consiste en contabilizar la presencia de los clústeres de emociones en el recurso analizado y escoger aquella emoción que tenga asociada mayor valor de ocurrencia.

A la ocurrencia de los clústeres de cada emoción se le asignará un peso que variará dependiendo si las palabras clave asociadas a las emociones aparecen de forma ‘normal’, intensificada o mitigada, asignando a cada forma los pesos 1, 1.5 y 0,5, respectivamente. Este sistema de pesos se implementa para identificar aquellos casos en que una emoción aparece con una intensidad mayor o menor, pudiendo darle, en consecuencia, mayor o menor valor en la determinación de la emoción del recurso. Por un lado, como ejemplo de ocurrencia de emoción en forma intensificada sería el caso del fragmento de texto ‘Muy alegre’ donde el adjetivo ‘alegre’ que se encuentra en el clúster <<*Alegría*>> se encuentra acompañado del adverbio cantidad ‘muy’ que se trata de un intensificador que indica un mayor nivel de dicha emoción. Por otro lado, el caso de los mitigadores se podría ejemplificar con el fragmento de texto ‘Poca tristeza’ donde el nombre ‘tristeza’, que corresponde al clúster <<*Tristeza*>>, está precedido por el determinante de cantidad ‘poco’ que indica un menor nivel de intensidad de dicha emoción.

Por último, una vez contabilizadas las ocurrencias de las emociones, se establece como emoción del recurso aquella con mayor peso asociado. En el caso que exista más de una emoción con el mismo peso se escoge como emoción resultante aquella que ha sido encontrada en primer lugar, basándose en el criterio de que la primera emoción expresada es la primaria.

Análisis de la emoción global

La detección de la emoción global asociada al evento seleccionado por el usuario, tanto para el contexto de jornada como para el semanal, se realiza una vez clasificados todos los recursos correspondientes a dichos contextos. Se calcula el porcentaje de *tweets* asociados a cada emoción y se proporciona como respuesta del presente módulo la emoción mayoritaria junto con los *tweets* asociados a dicha emoción. Tanto el resultado la emoción global extraída junto con los identificadores de los recursos asociados como los porcentajes correspondientes a cada emoción resultante se almacenan en base de datos como objetos ‘Observación’, con la estructura presentada en la dos siguientes figuras.

```
"_key": "708093040",
"_id": "observations/708093040",
"_rev": "_aXsyUC---",
"indicator": "3",
"metric": "301",
"perValue": null,
"socialNetwork": "TWITTER",
"time": 1587463049239,
"value": "Sevilla At.-28-Alegría-1234505475317739522-1236583768686768128-1236364944028971015-1234813271330308098"
```

Figura 22. Ejemplo de observación de emoción global en formato JSON

```
"_key": "708096413",
"_id": "observations/708096413",
"_rev": "_aXs2FRG---",
"indicator": "3",
"metric": "302",
"perValue": null,
"socialNetwork": "TWITTER",
"time": 1587463292931,
"value": "Villarrobledo-28-75;0;0;8;4;12"
```

Figura 23. Ejemplo de observación de porcentajes de emociones en formato JSON

Por otro lado, para el caso de análisis teniendo en cuenta el contexto objetivo, es decir, el resultado, una vez se ha calculado el porcentaje de recursos asociados a cada emoción, se ha aplicado a dicha probabilidad una función exponencial para así potenciar las emociones positivas en caso de resultado favorable o, en cambio, las emociones negativas ante un resultado desfavorable. La función exponencial se aplica de la siguiente forma:

$$f(x) = a^x \quad a = \text{Porcentaje de tweets asociado a emoción}$$

$$x = \text{Exponente asociado al resultado} \quad (\text{Eq. 1})$$

Los exponentes asociados al resultado para emociones positivas son los siguientes:

Resultado partido	Exponente asociado
Ganar	0,5
Empate	2
Perder	3,5

Tabla 48. Exponentes según resultado de función exponencial para emoción positiva.

Por otro lado, para el caso de emociones negativas los exponentes empleados en la función se muestran a continuación:

Resultado partido	Exponente asociado
Ganar	3,5
Empate	2
Perder	0,5

Tabla 49. Exponentes según resultado de función exponencial para emoción negativa.

Se ha escogido dicha función junto con los exponentes determinados para cada tipo de resultado porque permite obtener valores entre 0 y 1 para el caso de $0 < a < 1$ y $x > 0$. Una vez aplicada la función con el exponente correspondiente al tipo de emoción, como en los dos estrategias anteriores, se escoge la emoción con mayor porcentaje resultante. La aplicación de dicha función se utiliza como método para poder tener en cuenta el contexto

objetivo del evento en el análisis de la emoción global presente para los equipos y la jornada seleccionados.

5.5. Submódulo de Social Media News de muestra de resultados

El resultado obtenido del módulo de análisis será visualizado a través de la plataforma ya existente SMART-RTVE, por lo que el desarrollo de la interfaz se ha realizado mediante tecnologías que permiten la integración de la misma con dicha plataforma.

Tal como se ha adelantado en el capítulo 4.1 Infraestructura tecnológica, el desarrollo de la interfaz, tanto para la interacción del usuario mediante la selección de los parámetros a estudiar (equipo y jornada) como para la visualización de los resultados de la análisis, se ha llevado a cabo mediante el *framework* Angular. Para ello, se ha creado un nuevo componente en la plataforma SMART-RTVE llamada ‘News’ que incluye la totalidad de las funcionalidades del proyecto Social Media News. En este componente existe un contenedor que recoge la selección de parámetros y otro dos dedicados a la muestra de resultados, uno para los resultados del presente proyecto y otro para la parte de generación de noticias incluido en el proyecto Social Media News. Cabe destacar que para la creación de dichos contenedores y los diferentes elementos visuales que contiene la interfaz se ha utilizado la librería de estilos Angular Material [51].

En el contenedor situado en la parte superior de la web se sitúan los elementos necesarios para la selección por parte del usuario de los parámetros del sistema y la realización de las llamadas a los servicios web correspondientes. Como se muestra en la Figura 24, el usuario ha seleccionado uno de los equipos objeto de estudio, en concreto el Talavera de la Reina, y la jornada 28. Una vez seleccionados dichos parámetros, el sistema muestra el equipo contrincante mediante la llamada al servicio web de obtención de la información deportiva, indicando cuál de ellos atiende al rol de equipo local y visitante. Cabe destacar que el sistema establece por defecto como valor de jornada la última jornada ocurrida, sin embargo este

valor puede ser modificado por el usuario a través de la interfaz, como se ha comentado anteriormente.



Figura 24. Componente Interfaz para la selección de parámetros.

Por otro lado, en la Figura 25 se muestra el contenedor que contiene los elementos de visualización de los resultados obtenidos del sistema de análisis. En la parte izquierda se encuentra un gráfico de tipo araña que muestra el porcentaje de recursos asociados a cada categoría emocional analizada para las estrategias de contexto de jornada y de contexto semanal. Este elemento gráfico ha sido generado utilizando la librería Highcharts.js. Además, en el lado derecho se muestra la emoción global extraída para cada estrategia y el listado de *tweets* asociados a dicha emoción. Por último, en la parte superior se encuentra un elemento de tipo 'toggle' para la selección del equipo (local o visitante) del que se desea visualizar el resultado del análisis.



Figura 25. Componente Interfaz para la visualización de resultados.

6. Pruebas

En el presente capítulo, se especifica el plan de pruebas llevado a cabo para constatar el correcto funcionamiento del sistema. Por un lado, se comprueban aquellos aspectos recogidos en los requisitos (especificados en capítulo 3.2 Especificación de requisitos) mediante la realización de pruebas para su validación. Por otro lado, se evaluarán los resultados obtenidos en la clasificación emocional realizada por el sistema mediante el uso de diferentes métricas que permitirán medir su rendimiento.

6.1. Pruebas de validación

Para la especificación de las pruebas de validación realizadas se muestra, a continuación, la plantilla empleada para la documentación de las mismas y la justificación de los campos que contiene.

ID	
Requisitos implicados	
Descripción	
Acciones	
Resultado	

Tabla 50. Tabla ejemplo de definición de prueba de validación.

La definición de los campos es la siguiente:

- **ID:** identificador único de la prueba. Se trata de un código con el siguiente formato: PR– XY. Siendo XY un código numérico donde X e Y son números de 0 a 9.
- **Requisitos implicados:** enumeración de los identificadores de aquellos requisitos que son comprobados por la prueba definida.
- **Descripción:** explicación sobre la funcionalidad a probar.
- **Acciones:** pasos a llevar a cabo en la prueba.
- **Resultado:** respuesta obtenida tras la ejecución de la prueba.

A continuación, siguiendo el formato establecido, se presentan las pruebas realizadas:

ID	PR-01
Requisitos implicados	RF-01, RF-02, RNF-03, RNF-05, RNF-06, RNF-10, RNF-11, RNF-12
Descripción	Obtención y almacenamiento de la información deportiva (contexto objetivo) del servicio externo proveedor.
Acciones	<ol style="list-style-type: none"> 1. Lanzamiento de la tarea encargada de consumir el <i>feed</i> del proveedor de información deportiva y almacenamiento de dicha información en base de datos. 2. Acceso a base de datos.
Resultado	El sistema invoca correctamente el servicio externo proveedor de información deportiva y en base de datos se encuentra almacenada la información consultada.

Tabla 51. PR-01 Obtención y almacenamiento contexto objetivo

ID	PR-02
Requisitos implicados	RF-02, RF-03, RNF-03, RNF-06, RNF-10, RNF-12
Descripción	Consulta de información sobre evento deportivo (contexto objetivo) dado un equipo y jornada
Acciones	<ol style="list-style-type: none"> 1. Lanzamiento de la tarea encargada de consultar de base de datos la información deportiva relativa a un equipo y jornada.
Resultado	El sistema se conecta correctamente a la base de datos y obtiene un objeto que contiene la información deportiva consultada.

Tabla 52. PR-02 Consulta contexto objetivo.

ID	PR-03
Requisitos implicados	RF-04, RNF-03
Descripción	Monitorización de las cuentas y <i>hashtags</i> relacionados con los equipos de interés.
Acciones	<ol style="list-style-type: none"> 1. Definición de las cuentas y <i>hashtags</i> relacionados con equipos de fútbol de Segunda B a monitorizar en la plataforma SMART-RTVE . 2. Acceso a la plataforma SMART-RTVE. 3. Acceso a base de datos.
Resultado	La plataforma SMART-RTVE recoge correctamente los recursos asociados a las cuentas y <i>hashtags</i> definidos y estos se encuentran almacenados en base de datos.

Tabla 53. PR-03 Monitorización de cuentas y *hashtags*

ID	PR-04
Requisitos implicados	RF-05, RNF-03, RNF-06
Descripción	Consulta de los recursos asociados a un equipo y jornada.
Acciones	1. Lanzamiento de la tarea encargada de consultar de base de datos los recursos asociados a un equipo y jornada.
Resultado	El sistema se conecta correctamente a la base de datos y obtienen un listado de objetos que contiene la información relativa a las publicaciones asociadas al equipo y jornada especificados.

Tabla 54. PR-04 Consulta de recursos.

ID	PR-05
Requisitos implicados	RF-06, RF-07, RF-08, RF-10, RF-13, RNF-03, RNF-06, RNF-09, RNF-13
Descripción	Detección de la emoción de los recursos asociados a los equipos contrincantes y jornada especificados.
Acciones	1. Lanzamiento de la tarea encargada del análisis de emociones para los equipos contrincantes y jornada especificados. 2. Acceso a base de datos.
Resultado	El sistema devuelve la emoción global detectada para cada equipo y en base de datos se encuentran almacenadas las observaciones que recogen el resultado del análisis.

Tabla 55. PR-05 Detección de emoción de los recursos.

ID	PR-06
Requisitos implicados	RF-09, RNF-03, RNF-06, RNF-08.
Descripción	Conexión a Knowledge Manager (KM) y consulta de la ontología.
Acciones	1. Lanzamiento de la tarea encargada de establecer la conexión con la herramienta KM y usar la API Cake para la indexación de los recursos para la consulta de la ocurrencia de patrones y términos de la ontología.
Resultado	El sistema se ha conectado correctamente a la licencia de uso de KM y se ha obtenido un listado con los patrones y términos de la ontología detectados.

Tabla 56. PR-06. Consulta ontología.

ID	PR-07
Requisitos implicados	RF-11, RF-12, RF-13, RF-14, RF-15, RNF-01, RNF-02, RNF-03, RNF-06, RNF-07, RNF-09
Descripción	Consulta del resultado del análisis de emociones para el equipo y jornada seleccionados por el usuario.
Acciones	<ol style="list-style-type: none"> 1. Acceso a la interfaz del sistema. 2. Selección por parte del usuario de los equipos y jornada a consultar. 3. Llamada al servicio web de consulta de base de datos la emoción inferida para los equipos y jornada seleccionados.
Resultado	El sistema ha recogido parámetros seleccionados web y se ha invocado el servicio web correctamente, se conecta a base de datos y devuelve como respuesta la emoción global inferida y los <i>tweets</i> asociados obtenidos de base de datos para ambos equipos.

Tabla 57. PR-07 Consulta del resultado del análisis.

ID	PR-08
Requisitos implicados	RF-14 ,RF-15, RF-16, RNF-01, RNF-02, RNF-03, RNF-04, RNF-06.
Descripción	Visualización de los resultados.
Acciones	<ol style="list-style-type: none"> 1. Acceso a la interfaz del sistema. 2. Selección por parte del usuario de los equipos y jornada a consultar. 3. Llamada al servicio web de consulta de base de datos la emoción inferida para los equipos y jornada seleccionados.
Resultado	Se muestra correctamente mediante interfaz los resultados obtenidos del análisis de emociones para los parámetros seleccionados.

Tabla 58. PR-08 Visualización del resultado del análisis.

6.2. Evaluación de los resultados de la clasificación

Para evaluar los resultados obtenidos de la clasificación emocional de los recursos analizados se ha elaborado una muestra aleatoria de 250 *tweets* que han sido recogidos por la plataforma SMART-RTVE y analizados por el sistema, y se ha proporcionado la misma a tres anotadores para su clasificación manual dentro de las siguientes categorías: Neutro, Alegría, Tristeza, Ira, Sorpresa, Aversión y Miedo. En base a aquellos *tweets* que hayan sido etiquetados con

la misma categoría por los 3 anotadores, es decir, en los que estos hayan coincidido en la emoción detectada, se evaluará la clasificación del sistema mediante el cálculo de la exactitud y de las métricas para la evaluación del rendimiento precisión, exhaustividad y valor-F.

Cabe destacar que el uso de la métrica *accuracy* (exactitud) es altamente utilizada para evaluar los problemas de clasificación, sin embargo en este caso dicha métrica no resulta del todo adecuada para una correcta evaluación del sistema ya que la muestra empleada no está balanceada, es decir, el número de observaciones de cada clase no está igualado [52].

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (\text{Eq. 2})$$

Por tanto, a la hora de evaluar el sistema de clasificación se tendrá en cuenta también el valor de las métricas precisión, exhaustividad y valor-F.

En primer lugar, la precisión se trata de una métrica que se refiere a la calidad de la clasificación realizada, es decir, indica que porcentaje de elementos etiquetados con una categoría pertenecen realmente a dicha categoría. Se calcula dividiendo el número de verdaderos positivos entre la suma entre verdaderos positivos y falsos positivos.

$$precision = \frac{TP}{TP + FP} \quad (\text{Eq. 3})$$

En segundo lugar, la exhaustividad (*recall*) consiste en una métrica que mide la cantidad de elementos de una determinada categoría que el sistema tiene la capacidad de identificar. Su cálculo se realiza mediante la división del número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos.

$$recall = \frac{TP}{TP + FN} \quad (\text{Eq. 4})$$

Por último, el valor-F o F1 se trata de una métrica que permite la combinación de los valores de precisión y exhaustividad en un solo indicador.

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (\text{Eq. 5})$$

El resultado obtenido del cálculo de la exactitud es de un 65%. Por otro lado, se presentan a continuación el valor de las métricas precisión, exhaustividad y valor-F para cada una de las clases a detectar y el valor medio de F1 (*macro-F1*):

Categoría	Precisión	Recall	F1
Neutro	0,96	0,56	0,71
Alegría	0,79	0,76	0,78
Tristeza	0,3	0,75	0,43
Sorpresa	0,75	1	0,86
Ira	0,63	0,56	0,59
Aversión	0,33	0,67	0,44
Miedo	0	0	0
Macro-F1	0,54		

Tabla 59. Cálculo métricas primera evaluación de clasificación

Antes de analizar el valor de la métricas obtenido para cada categoría, cabe destacar el bajo porcentaje de coincidencia entre evaluadores (un 35%) lo que señala la alta complejidad de este tipo de problema de clasificación debido a su alta subjetividad.

A continuación, se presentan los principales aspectos y conclusiones a destacar tras la obtención del valor de las métricas anteriores. En primer lugar cabe destacar el hecho de que los evaluadores no han empleado la categoría ‘Miedo’ para etiquetar los *tweets* de la muestra, lo que indica que dicha clase puede que no se ajuste al campo en el que se realiza el análisis. Por tanto, se podría tomar la decisión de eliminar dicha categoría del problema de clasificación para la mejora del sistema. El valor *macro-F1* obtenido de la misma muestra si se prescinde de dicha categoría es de un 63% lo que supone una mejora del rendimiento de un 9%.

En segundo lugar, otra problemática encontrada es el bajo valor de precisión en ‘Aversión’ y ‘Tristeza’, a pesar del alto valor en el *recall*, lo que indica que la calidad de la clasificación para dichas clases es baja. Tras el análisis de este aspecto, se ha llegado a la conclusión del que sistema presenta ciertas dificultades para diferenciar dichas emociones respecto al conjunto de emociones negativas estudiadas (‘Tristeza’, ‘Ira’ y ‘Aversión’).

Por otro lado, debido al bajo porcentaje de coincidencia presente entre evaluadores, se ha realizado una segunda evaluación en la que se empleará la misma metodología sin embargo se considerará como acierto o verdadero positivo el caso en el que el sistema coincide en su clasificación con al menos uno de los evaluadores. Además se ha prescindido de la emoción ‘Miedo’. Los resultados obtenidos de esta segunda evaluación se presentan en la siguiente tabla:

Categoría	Precisión	Recall	F1
Neutro	0,98	0,69	0,81
Alegría	0,82	0,88	0,85
Tristeza	0,62	0,97	0,76
Sorpresa	0,55	1	0,71
Ira	0,63	0,83	0,71
Aversión	0,27	0,8	0,4
Macro-F1	0,71		

Tabla 60. Cálculo métricas segunda evaluación de clasificación

En este caso, observamos que se ha obtenido una media del valor-F del 0,71, lo que indica una buen rendimiento del sistema. Sin embargo, encontramos la misma problemática con la clase ‘Aversión’ que en la evaluación anterior, el valor de precisión es muy bajo, lo que indica que quizás sería adecuado prescindir de dicha clase por su confusión con el resto de las emociones negativas del sistema o bien modificar la ontología para la mejora de resultados para dicha emoción.

Por último, se muestran algunos ejemplos en los que el sistema ha realizado una correcta clasificación, es decir, coincidente con la categoría etiquetada por los evaluadores. Además, se indica aquellas palabras clave o patrones que han determinado la emoción detectada:


Tweet	Categoría
 Estadio Nuevo Vivero listo para el encuentro entre el @CDBadajoz y el @CFTalavera_ a las 17:00 h. ⚽ #Royalverd #2bg4 #Badajoz https://t.co/VnJNBxxAMB	Neutro
FINAL 0-2 Se acabóooo. Importantísima victoria del #RecreativoGranada en Villarrobledo. ¡¡Partidazo de todo el equipo. VAMOS!! 🙌🙌🙌 #VillarrobledoRecreativoGranada #CanteraNazari	Alegria
Lástima el resultado @maurolucero96 pero se ha competido y luchado hasta el final.	Tristeza
@AlbertMosca @formacvillarrub Para mí si es sorpresa, pero aun así la temporada del Villarrubia es para quitarse el sombrero.	Sorpresa
Poco que decir, si sales así a un partido donde te juegas la vida lo que te puede pasar es esto, un gol en el 92 tras quedarnos con 9 jugadores no es excusa, hemos hecho un partido de MIERDA, ALCOY VETE YA https://t.co/nFPSeunLtv	Ira
Se tuvo muy poca memoria con Dani Pino el año pasado, salvó a los chavales y al año siguiente se le echó por no ir 1º!! Y me tengo que creer que si méritos que si pollas etc, cada vez me da más ASCO todo, está todo PODRIDO, la meritocracia no existe ni en el club ni en la ciudad	Aversión

Tabla 61. Ejemplos resultado de la clasificación de emociones

7. Marco Legislativo

En primer lugar, cabe tener en cuenta la licencia de uso de los datos utilizados para la construcción de la ontología que conforma la base de conocimiento. El léxico de emociones SEL cuenta con un acuerdo de licencia, especificado en la web de descarga [53], en el que se establece el libre uso de dicho recurso en el ámbito académico bajo el requisito de citar los artículos referentes al mismo, además se indica que para el uso comercial se requiere la consulta previa con el autor. Por otro lado, los *emojis* obtenidos del trabajo “Emoji as Emotion Tags for Tweets” cuentan con un licenciamiento también de libre uso académico con el requisito de referenciar dicha obra.

En segundo lugar, para el uso de la herramienta Knowledge Manager se ha requerido de la licencia proporcionada por la compañía propietaria del software *The Reuse Company*, que ofrece dicha licencia de uso para la realización de los proyectos de la Cátedra RTVE-UC3M.

Por otro lado, cabe destacar la normativa aplicable al tratamiento de los datos personales obtenidos de Twitter, entre la que encontramos el Reglamento (UE) 2016/679 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación (RGPD) [54]. y su aplicación en el territorio español mediante la Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD) [55] que regula el tratamiento de datos personales para la protección de las personas. En este caso los datos personales tratados son obtenidos a través la plataforma SMART-RTVE que emplea la API de Twitter, la cual se encarga de obtener el consentimiento por parte de los usuarios de la recogida de datos a través de la política de privacidad aceptada por los mismos [56]

Además, en la visualización mediante interfaz gráfica de la publicaciones analizadas se debe tener en cuenta la Ley de Propiedad Intelectual [57] donde, de acuerdo al artículo 1, se especifica que “La propiedad intelectual de una obra literaria, artística o científica corresponde al autor por el solo hecho de su creación”, por lo que el autor de un *tweet* posee los derechos de dicho contenido. Por tanto, para el uso de este recurso es necesario que se cite el mismo [58], en este caso indicando el nombre de usuario del autor y el enlace al *tweet*.

8. Entorno Socioeconómico

El entorno en el que se encuadra el presente proyecto es en el de la digitalización de los medios de comunicación. Hoy en día el foco de actuación del sector está dirigido hacia los lectores, siguiendo la tendencia denominada ‘*customer media*’ [59], como resultado de una sociedad en la que la forma de consumir la información ha cambiado en un alto grado. Además, se suma la cuarta revolución industrial vivida en la economía española donde el análisis de datos personales es uno de sus principales motores y en el que las tecnologías Big Data, Data Science, Inteligencia Artificial, etc. juegan un gran papel.

Además, como consecuencia del proceso de transformación, la cadena de valor del sector también se ha transformado y esto a su vez supone un cambio en su modelo de negocio. Anteriormente los medios de comunicación controlaban toda la cadena de valor, sin embargo con la distribución de la información a través de la aparición de nuevas plataformas el control ejercido por dichos medios es cada vez menor. Como claro ejemplo de la importancia del proceso de digitalización de los medios de comunicación y la necesidad de adaptación de su modelo de negocio se podría destacar el hecho de que el 40% de los ingresos publicitarios tienen como origen recursos de publicidad *online* [60].

Una de las tendencias a destacar que resulta uno de los puntos clave para llevar a cabo el enfoque hacia el cliente mencionado por parte de los medios es el uso de la web 2.0, en la que el protagonista es la interacción con el usuario. Este tipo de web ha ido evolucionando, llegando a surgir las redes sociales u otros portales donde el usuario es el encargado de crear y compartir contenido. Además, en esta evolución se incorporan distintas tecnologías en las que se emplean bases de datos, *cookies*, Big Data y otras herramientas de marketing. Bajo esta tecnología se encuentra la plataforma SMART-RTVE, con la que se integra el presente proyecto.

Por otra parte, otra de las grandes tendencias empleadas en este proceso de transformación es la Inteligencia Artificial, en concreto resaltar el uso de tecnologías de procesamiento del lenguaje natural. El uso de esta tecnología resulta esencial para el tratamiento del texto no estructurado presente en las publicaciones de redes sociales, blogs, etc. Mediante las diferentes técnicas de NLP es posible la extracción de conceptos, identificar relaciones entre

distintos elementos o, incluso, cómo se pretende en el sistema propuesto, la detección de emociones. Por tanto, el uso de dicha tecnología permite la extracción de información de un gran número de fuentes de forma automática y a menor coste [61].

En base a esta situación del entorno actual de los medios de comunicación, el presente proyecto busca, mediante el uso de los medios digitales, en concreto de las redes sociales, la obtención de información de alto valor respecto al comportamiento y las opiniones de sus usuarios. Dicha información resulta de gran utilidad en el sector mencionado para un mayor conocimiento del cliente o usuario, aunque dicho beneficio podría ser aplicable a otros sectores. Además, estos datos de gran valor son obtenidos a un menor coste económico, de esfuerzo y de tiempo requerido respecto a otros métodos como pueden ser encuestas públicas de gran escala o cuestionarios.

En el caso específico del sistema propuesto, su finalidad es el análisis de un público actualmente no alcanzable para RTVE para así generar contenido personalizado en base a la información obtenida. Esto supone un impacto positivo directo en este medio de comunicación ya que esta mayor orientación hacia el usuario consigue una mayor atracción y satisfacción de este. En los medios digitales esto se traduce en un mayor número de visitas, mejora del posicionamiento en buscadores, etc.

9. Planificación

A continuación, se detalla la planificación seguida para el desarrollo del sistema propuesto. En primer lugar, se especifican las fases propias del proyecto y las fechas de consecución de las mismas, así como la asignación de recursos en cada una de ellas. Posteriormente, para una representación clara de dicha planificación se muestra un Diagrama de Gantt con las diferentes tareas de las fases de proyecto ubicadas en el tiempo de realización

9.1. Fases del proyecto

En primer lugar, cabe señalar que el inicio del proyecto tiene como fecha de inicio el 1 de noviembre de 2019 y finaliza el 1 de junio de 2020, por lo que la duración es de 7 meses. La realización de la planificación se basa en el calendario laboral de la Comunidad de Madrid y se establece una dedicación semanal de 20 horas.

En el presente proyecto se distinguen cinco grandes fases: análisis del problema, análisis del sistema, diseño de la solución, implementación y pruebas. En estas fases se distinguen a su vez diferentes tareas que se encuentran especificadas en la Tabla 62. De forma paralela a las diferentes fases se realiza la documentación del proyecto.

Nombre de tarea	Duración	Comienzo	Fin	Asignación Jefe de proyecto	Asignación Analista
Propuesta del proyecto	20 horas	vie 01/11/19	vie 08/11/19	1	0
Definición del problema	20 horas	vie 08/11/19	vie 15/11/19	0,8	0,2
Estado del arte	40 horas	vie 15/11/19	vie 29/11/19	0,25	1
Análisis	60 horas	vie 29/11/19	vie 20/12/19	0,3	1
Especificación de requisitos	40 horas	vie 29/11/19	vie 13/12/19		
Definición casos de uso	20 horas	vie 13/12/19	vie 20/12/19		
Diseño	60 horas	vie 20/12/19	vie 10/01/20	0,3	1
Arquitectura a alto nivel del sistema	16 horas	vie 20/12/19	jue 26/12/19		
Vista lógica	8 horas	jue 26/12/19	lun 30/12/19		
Vista de desarrollo	16 horas	lun 30/12/19	vie 03/01/20		
Proceso de detección de emociones en los tweets	20 horas	vie 03/01/20	vie 10/01/20		
Implementación	324 horas	vie 10/01/20	lun 04/05/20	0	1
Módulo de obtención de recursos	40 horas	vie 10/01/20	vie 24/01/20		
Módulo de configuración y construcción de la ontología	80 horas	vie 24/01/20	vie 21/02/20		
Módulo obtención contexto objetivo	12 horas	vie 21/02/20	mié 26/02/20		
Módulo de análisis y clasificación de emociones	128 horas	mié 26/02/20	vie 10/04/20		
Submódulo de Social Media News de muestra de resultados	64 horas	vie 10/04/20	lun 04/05/20		
Pruebas	80 horas	lun 04/05/20	lun 01/06/20	0,5	1
Diseño plan de pruebas	32 horas	lun 04/05/20	jue 14/05/20		
Ejecución de pruebas	48 horas	jue 14/05/20	lun 01/06/20		

Tabla 62. Planificación de fases y asignación de recursos del proyecto.

9.2. Diagrama de Gantt

A continuación, se presenta el diagrama de Gantt que representa de forma gráfica la planificación de las diferentes etapas del proyecto situando las mismas en el tiempo e indicando sus dependencias.

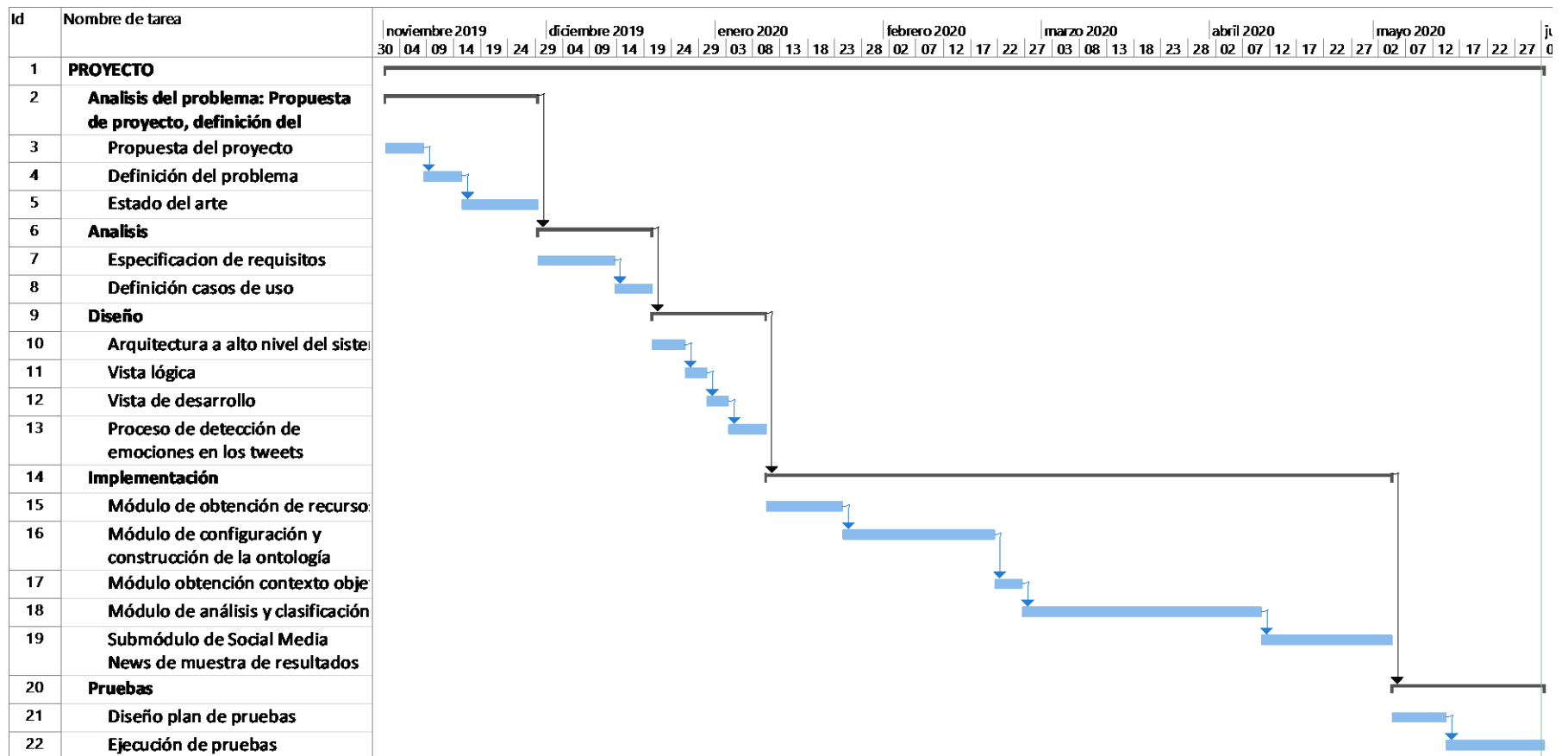


Figura 26. Diagrama de Gantt

10. Presupuesto

En presente capítulo se presenta la estimación del presupuesto del proyecto. En este se incluyen tanto los costes de personal, del material software y hardware, y los costes indirectos necesarios para el desarrollo del proyecto.

10.1. Coste de personal

El presente proyecto ha sido realizado por dos personas, una encargada del rol de jefe de proyecto y otra responsable de las tareas de analista y desarrollador. Para el cálculo de los costes incurridos en personal se debe tener en cuenta, en base a la planificación presentada en el Capítulo 9. Planificación, el número de horas imputables a las diferentes tareas proyecto.

El salario establecido se ajusta a la media salarial proporcionada por el portal de empleo Indeed basada en el estudio de los salarios extraídos directamente de fuentes conformadas por usuarios, ofertas de empleo publicadas y empresas. Para el caso del rol de jefe de proyecto la media salarial anual obtenida es de 36.930 € [62] y, por otro lado, el salario medio anual para un puesto de analista programador ASP.NET es de 24.567 € [63]. Cabe destacar que el valor del salario indicado es bruto y a este se debe añadir el valor el coste de la Seguridad Social soportado (un 30% sobre el sueldo base) que incluye los siguientes costes: contingencias comunes, tipo general de desempleo para un contrato indefinido, Fondo de Garantía Salarial y formación profesional. Además, para el cálculo de coste por hora realizado a continuación, se supone que la media salarial de referencia se encuentra asociada a una jornada laboral de 8 horas diarias durante los 20 días hábiles del mes.

Costes de personal		
Datos del empleado		
Apellidos	Álvaro Rodríguez	Sánchez Muñoz
Nombre	Jose María	Andrea
Puesto	Jefe de proyecto	Analista y Desarrollador
Jornada Laboral	2	4
Hora semanales	10	20
Retribución fija		
Salario Base/Año	36.930,00 €	24.567,00 €
Salario Base/Mes	3.077,50 €	2.047,25 €
Seguridad social		
Seguridad social/ anual	11.079,00 €	7.370,10 €
Seguridad social/ mes	923,25 €	614,18 €
Coste total de personal		
Anual	48.009,00 €	31.937,10 €
Mensual	4.000,75 €	2.661,43 €
Coste Hora Trabajada	25,00 €	16,63 €

Tabla 63. Cálculo coste personal estimado por hora trabajada

Una vez extraído el valor de coste por hora trabajada para cada rol de empleado, se calcula el coste de personal imputado al presente proyecto para estimar el gasto total en personal del mismo. Para ello se establece el número de horas invertidas por cada trabajador en las diferentes actividades que conforma el proyecto.

Costes de personal imputadas al proyecto			
Datos del empleado			
Apellidos	Álvaro Rodríguez	Sánchez Muñoz	
Nombre	Jose María	Andrea	
Puesto	Jefe de proyecto	Analista y Desarrollador	
Horas imputadas al proyecto por actividad			Total
Propuesta del proyecto	20	0	20
Definición del problema	16	4	20
Estado del arte	10	40	50
Análisis	20	60	80
Diseño	20	60	80
Implementación	0	324	324
Pruebas	40	80	120
Total horas imputadas	126	568	694
Coste Hora Trabajada	25,00 €	16,63 €	
Coste Total	3.150,59 €	9.448,06 €	12.598,65 €

Tabla 64. Cálculo del coste total estimado de personal

10.2. Coste de material

A continuación, se presentan los costes de material, tanto hardware como software, incurridos para el desarrollo del proyecto. En el coste de material también se incluye el coste en material fungible, pero en este caso por su escasez y bajo valor resulta despreciable.

10.2.1. Costes de Hardware

El coste de material hardware imputable al proyecto, teniendo en cuenta que para el cálculo de los costes de amortización la duración de proyecto es la establecida en la planificación (7 meses), se estima en la siguiente tabla.

Costes material Hardware					
Producto	Modelo	Unidades	Precio sin IVA	Vida útil	Coste imputable al proyecto
Ordenador portátil	Acer Swift 3 14"	1	577,69 €	24 meses	168,49 €
Ordenador sobremesa	Packard Bell Imedia S2883	1	355,85 €	24 meses	103,79 €
Monitor	Philips 243V7QDSB 23.8"	2	82,64 €	24 meses	48,21 €
TOTAL					320,49 €

Tabla 65. Costes estimados de material hardware

10.2.2. Costes de Software

El coste asociado al material software ha resultado nulo debido al ámbito académico en el que se encuentra el presente proyecto, por lo que las licencias de los diferentes productos software utilizados son gratuitas.

Costes material Software			
Producto	Unidades	Precio sin IVA	Coste Total
Visual Studio 2019	1	0 €	0 €
Knowledge Manager	1	0 €	0 €
Paquete Office	1	0 €	0 €
Windows 10 Education	1	0 €	0 €
TOTAL			0 €

Tabla 66. Costes estimados de material software

10.3. Costes Indirectos

Los costes indirectos asumidos se establecen como 15% adicional sobre el valor de los costes directos y estos incluyen los costes asociados a reuniones, desplazamientos, oficina OTRI y dietas.

10.4. Coste Total del Proyecto

Por último, se presenta a continuación el presupuesto total del proyecto, donde se integran los costes directos e indirectos, los márgenes de seguridad y beneficio, y los impuestos a considerar.

Coste Total Proyecto	
Coste Personal	12.598,65 €
Coste Material	320,49 €
Costes Directos	12.919,14 €
Costes Indirectos (15%)	1.937,87 €
Total SIN IVA	14.857,01 €
Margen seguridad(10%)	1.485,70 €
Margen Beneficio (15%)	2.228,55 €
IVA (21%)	3.899,96 €
TOTAL	22.471,23 €

Tabla 67. Presupuesto estimado del proyecto.

Por tanto, en base al presupuesto estimado del presente proyecto, el coste total del mismo asciende a **VEINTIDOS MIL CUATROCIENTOS SETENTA Y UN EUROS CON VEINTITRES CÉNTIMOS** (22.471,23 €), IVA incluido. La validez de dicho presupuesto tiene plazo de 30 días desde la presentación del proyecto.

11. Conclusiones y Trabajo Futuro

11.1. Conclusiones

Tras el diseño y desarrollo del sistema así como su posterior evaluación, se presentan a continuación la conclusiones obtenidas en base al cumplimiento de los objetivos marcados por el mismo.

En primer lugar, se han estudiado y valorado los diferentes enfoques y recursos existentes para la resolución del problema de ingeniería a abordar en el proyecto, la detección automática de emociones en texto, y se han detectado ciertos aspectos que han supuesto varias limitaciones. Por un lado, una falta del desarrollo de dicho campo respecto al existente para el análisis de sentimientos o polaridad y, por otro lado, también destaca la escasez de recursos para el análisis disponibles en español (*datasets*, léxicos, etc.). Pese a dichas problemáticas, se ha propuesto una solución efectiva para el problema que consiste en un sistema basado en conocimiento a partir de una ontología conformada por el léxico SEL y una serie de *emojis*, que permite la detección de las categorías de emoción definidas por Ekman en textos procedentes de redes sociales.

Tras la definición de la solución, se ha llevado a cabo el análisis, diseño e implementación de la misma. En esta se ha desarrollado un sistema de detección automática de emociones de comunidades deportivas en el que se han identificado las dimensiones de interés para el análisis (evento deportivo, liga, grupo, equipo, fecha del evento, resultado y emoción), se ha definido un proceso para el análisis de la gran cantidad de recursos procedentes de Twitter dónde se han aplicado técnicas de encaje de patrones y algoritmos para el filtrado de recursos relevantes y detección de emociones (tanto a nivel de *tweet* como global). Además, se emplea una arquitectura basada patrón Modelo-Vista-Controlador (MVC) adecuada para la implementación de la solución mediante servicios web de tipo REST y su integración con el sistema ya existente de SMART-RTVE. Por otro lado, se ha implementado una interfaz de usuario que cuenta con la validación del personal de RTVE y que permite la visualización de los resultados del análisis así como la interacción del usuario para la selección de parámetros.

Por último, además de la comprobación del correcto funcionamiento del sistema mediante la realización de pruebas de validación de requisitos, en base a los valores obtenidos de la exactitud y métrica F1, se ha determinado que el sistema cuenta con un rendimiento medio superior al 50%, con un valor de *accuracy* del 65% y un valor-F del 54%. Cabe destacar que el valor obtenido de F1 en el caso de prescindir de la clase de ‘Miedo’ por falta de relevancia en ámbito aplicado sería del 63% (mejorando el rendimiento en un 9 %). El sistema presenta mayor rendimiento en el caso de las emociones positivas, ‘Sorpresa’ con un 85% y ‘Alegría’ con 78%, y en el caso de la categoría ‘Neutro’ con un 71%. Sin embargo, dicho valor de rendimiento es mejorable y, por tanto, quedan abiertas nuevas líneas de mejora que se tratarán en el siguiente apartado 11.2 STrabajo Futuro.

11.2. STrabajo Futuro

Por último, para la mejora del sistema actual se especifican a continuación ciertas tareas cuya realización sería de interés para un trabajo futuro:

- Mejora de la ontología empleada, aspecto que resulta de dificultad debido a la falta de recursos para el análisis de emociones en texto en español, sin embargo sería beneficioso para la mejora del rendimiento de la solución.
- Investigación en la definición de patrones para la detección de emociones que no únicamente se basen en *keywords* que expresen la emoción de forma explícita sino que también recojan la expresión de la emoción de forma implícita.
- Redefinición de las categorías emocionales a clasificar para que se adapten en mayor grado al dominio en el que se aplica el análisis. Para ello sería necesario el estudio de otras clases de emoción a incluir o diferentes intensidades de las mismas, por ejemplo las emociones de la rueda de emociones de Plutchik.
- Inclusión de un mayor número de aspectos o factores para la consideración del contexto objetivo a la hora de determinar la emoción global de un equipo para una jornada (actualmente solo se considera el resultado), por ejemplo posición en la clasificación de la liga, equipos tradicionalmente enfrentados o enemigos, situación de ascenso o descenso, número de partidos consecutivos ganados o perdidos, etc.

- Estudio de desarrollo de un enfoque híbrido para la detección de emociones mediante la inclusión de aprendizaje automático que permitan una mejora del rendimiento actual de la clasificación. Para ello sería necesario obtener un *dataset* en español de gran tamaño y calidad, aspecto de dificultad ante la falta de recursos mencionada. Alto interés en el uso de SVM o redes de neurona recurrentes LSTM debido a los buenos resultados que ofrecen en el campo de detección de emociones en texto.
- Aplicación del sistema en otros ámbitos que puedan resultar de interés para RTVE y estudio de los resultados obtenidos para evaluar la capacidad de reutilización y el rendimiento de la solución en diferentes campos.

12. Extended Abstract

12.1. Abstract

The boom of social networks today has placed them as one of the main means of communication where users express their thoughts, sentiments or opinions and interact with other users and communities. One of the areas that generates a high level of emotional activity in social networks is sports, so this means is a source of valuable information to know the opinion or emotional state of their followers or supporters. With the aim of taking advantage of this information arises the objective of this project, which consists of the development of a system for automatic detection of the emotion expressed in social networks of certain communities or supporters in the course of sports events through the analysis of their publications.

This project emerges within the collaboration agreement between RTVE and UC3M, which conforms the RTVE-UC3M chair and whose purpose is the research in the analysis of social networks and its application in the audio-visual sector.

The proposed system aims to detect the emotion expressed by those followers and communities of interest, whose accounts are monitored, in order to offer this information to RTVE as support for decision making or personalization of content. The system, despite being designed to be applied to different fields, has as its first scope the analysis of the emotion associated to sports events in the 2B IV group football league, since these events are not included in the sports information currently provided by RTVE.

This document describes the analysis, design, implementation and evaluation of the proposed system, which uses a knowledge-based approach to the analysis of emotions, consisting of an ontology and a set of patterns, and provides to the user an interface for the selection of the team and the journey to be analysed, as well as for graphically visualising the global emotion inferred and the publications associated with this emotion.

12.2. Introduction

Social networks have become one of the main means of mass communication today, being in a mature stage of the life cycle. Through the social media, people are able to satisfy one of their basic necessities, which is to communicate, expressing their thoughts, feelings or opinions and interacting with other users and communities. For this reason, this means generates a large number of publications that are a source of valuable information for its study in many areas of knowledge, such as psychology, sociology, politics, marketing, computer science, etc.

From an analytical point of view, one of the fields that obtains a great profit from this large amount of information, in the form of publications and/or comments, is the analysis of sentiment, opinions or emotions. The use of publicly available data to carry out emotional analysis significantly reduces the costs, efforts and time required to manage large-scale public surveys and questionnaires.

To make possible the use of the large volume of information generated by social media, it requires tools for managing and processing great amounts of data automatically. Big Data technologies play an important role in this requirement and their growth is unstoppable in an age where digital transformation is the protagonist.

The applications of sentiment and emotion analysis in social networks are multiple. For example, in the field of marketing, the opinion obtained from users regarding a product or brand is used to design personalized marketing campaigns, to improve customer communication or to define retention strategies.

Given the SMART-RTVE platform, developed by the RTVE-UC3M chair, which allows the monitoring of accounts and *hashtags* on Twitter to obtain their publications, this project aims to cover the requirement of RTVE for analysing resources published by certain users and/or communities to find out their emotions or opinions regarding different topics, events, etc. In addition, the proposed system will show the result of the analysis in a visual way within the mentioned platform. The emotions or opinions extracted from the analysis will be used by

RTVE for decision making, such as the personalization of news content according to these parameters.

Specifically, this project aims to obtain the emotions present in football fans which are associated with matches played. Therefore, the system will focus on the detection of emotions of the resources published during the pre-match and the course of the match by the accounts and *hashtags* of teams, fans and media.

However, the proposed system is not intended to resolve this need of analysis only for the sports events mentioned, but rather the proposal seeks a general design that can be applied to other types of sports events or even other fields.

12.3. Motivation

Sports events are a topic that usually generates debate and in the sports information which is published in the media it is possible to perceive this controversy. Nowadays, there are many open data portals, like LFP or NBA.com, from which a lot of sports information is obtained, even in some cases with a certain critical nature. However, social networks, through the publication of comments or chronicles of sports events, are one of the media where this debate and the emotions expressed by the different communities in response to a certain result are best observed.

Social media plays an important role in football matches [2], especially Twitter, where a "parallel game" takes place. Before an important match (Champions, the classic, etc.), fans comment, debate and show their emotions and opinions, making this event a trend on Twitter.

Thus, there is clear evidence that a large amount of updated sports information with a high emotional nature is obtained from social networks. Through an analysis of sentiment, opinion or emotion of these data it is possible to obtain valuable information, which is highly representative for the strategic study of social behaviour.

12.4. Objectives

According to the requirements mentioned above, the overall goal of this project is to design and implement a system for the automatic detection of the emotions present among the

supporters during the course of sports events, through the analysis of their publications on social networks. This will be done using natural language processing and pattern matching techniques, the result of which (the inferred emotion) will be provided to the user in a visual way for its later use by RTVE staff for decision making and consequent actions.

In order to reach this objective, the specific objectives listed below must be achieved:

- Study of the methodologies and resources for the automatic detection of emotions and the existing emotional models for the definition of the categories on which the classification of resources will be based.
- Analysis of the proposed system, which specifies the different functionalities it has to perform, the interaction of users with it and the restrictions it must comply with.
- Definition of a methodology for the processing and integration of the resources related to the information of interest from Twitter.
- Determination of the set of dimensions to be identified in the sports information required to perform the analysis of emotions: sports event, league, group, team, event date, result, etc.
- Design of an appropriate system architecture that allows its future integration and reuse.
- Configuration of a suitable ontology and semantic patterns for the correct recognition of emotions and decision rules/algorithms to determine the emotion (or its intensity) according to the subjective and objective context.
- Integration of the social media monitoring platform SMART-RTVE and design and implementation of the service for querying the data to be analysed.
- Implementation of the analysis service and deployment of the algorithms for classifying the resources.
- Implementation of a graphic interface for user interaction and display of results.

- Testing to verify and validate the proper functioning of the system and evaluate the results of the classification.

12.5. Emotion detection

In the domain of computational linguistics, emotion analysis consists in the process of classifying a text according to the emotion it expresses. It can be understood as a study product of the natural evolution of the analysis of sentiments, which classifies a text in terms of the polarity of the subjective charge it presents (positive, negative or neutral).

It should be noted that the number of studies and methodologies present in the field of sentiment analysis are far superior to that of emotion detection, which has refined models and highly satisfactory results. This is an established field in the area of natural language processing, which is particularly successful in applications such as advertising [4], abstracts [5] and in improving the process of extracting information from texts [6], among others.

Although there is a long way to go for the emotion analysis to reach the results of the sentiment analysis, the additional information that it provides can mean a high improvement in the applications of the latter. For example, there are emotions that, despite having the same polarity, such as fear and anger (negative polarity), can influence the state or behaviour of users in different ways, so that their distinction is very valuable for application in fields such as marketing or socio-political.

12.5.1. Methodologies for emotion detection

Among the methodologies used in the detection of emotions in texts, specifically in texts from social networks, three approaches can be differentiated [29]: based on keywords or lexicons, machine learning based and hybrid.

The keyword-based method is a simple and intuitive approach that aims to find patterns in the analysed text that fit with keywords associated with emotions. To do this, it is necessary to split the text to be analysed into words and determine the part of speech of each of them, identifying whether they are adjectives, nouns, verbs or adverbs, which are the grammatical categories most likely to be emotionally charged. Once this labelling has been performed,

the existence of keywords associated with emotions is checked within the set of labelled words. The emotion that matches with the analysed sentence will be the emotion assigned to that sentence.

The Lexicon based method is similar to the previous one but uses emotion lexicons instead of a list of keywords for the detection of emotions.

Between these two methods, works based on lexicons of emotions are more popular. For example, Badugu and Suhasini [30] propose a system with a rule-based approach that classifies tweets within a set of emotional categories. This classification is based on Russell's Circumplex Model of Affect [12] and the lexicon used for the knowledge base is Sentiwordnet 3.0 [31]. Although this word dictionary is for sentiment analysis, the polarity scores of each word in the lexicon have been used to classify the words according to the four categories mentioned.

The other most popular approach to emotion detection in text is the use of machine learning techniques, both supervised and unsupervised. This method allows the automatic classification of texts from the design of a model to train the classifier through a significant number of examples that conform the training data. In this way, it is possible to classify automatically the rest of the data (learning through an algorithm that allows the identification of patterns or features).

One study that stands out within this approach is that of Shah, Reyadh, Shaaf, Ahmed and Sithilen [32] who have worked on an emotion detection system in tweets on the IWA-2018 dataset [33]. The proposed model is based on the lexicons WordNet-Affect and EmoSenticNet [34] with the use of supervised classifiers (Naïve Bayesian, Decision Trees and Support Vector Machines). In this study the highest accuracy has been achieved by using the EmoSenticNet lexicon and the Support Vector Machine algorithm for training.

Finally, there is the use of a hybrid approach. This type of method combines the two previous approaches and can take advantage of the benefits of both approaches and achieve greater precision. An example of this type of approach is the work "Emotion Recognition from Text Using Semantic Labels and Separable Mixture Models," [36] which describes a hybrid model that uses a Chinese lexicon and rule system for the extraction of emotion-related features and

semantics and an automatic learning-based classifier that uses SVM, Naïve Bayesian and Max Entropy algorithms.

12.6. Proposed system

The main problem presented in this project is to obtain knowledge to decision support through techniques of analysis of short text fragments in natural language (tweets). These texts are obtained by the existing platform SMART-RTVE, which allows the extraction from social networks of a large amount of resources, specifically from the social network Twitter.

The proposed solution is a knowledge-based system that uses natural language processing and a knowledge base from which to apply rules or algorithms for the inference of emotions in text. Regarding the knowledge on which the system will be based, it is composed of the 'SEL' lexicon, which is one of the few available lexicons of emotions in Spanish, and the set of emojis classified by Wood and Ruder's study, both based on the model of the 6 basic emotions of Ekman. The solution uses the 'Knowledge Manager' tool, which allows both the management of the knowledge on which the system is based and the application of the natural language processing techniques for text analysis.

Finally, it should be mentioned that the information relating to the sports events analysed is obtained through an external sports information provider, which has a contractual relationship with RTVE, and the result of the analysis is provided to the user for its visualization through a user interface.

12.6.1. Emotion detection process in tweets

As stated above, the designed analysis system has a knowledge-based approach in which two parts can be distinguished: a knowledge base for emotion recognition, which is composed of the SEL emotion lexicon, a group of emojis tagged with emotions defined in "Emoji as Emotion Tags for Tweets" and a set of patterns, and, on the other hand, an inference engine that is based on rules. The design of the emotion recognition process in tweets is shown below:

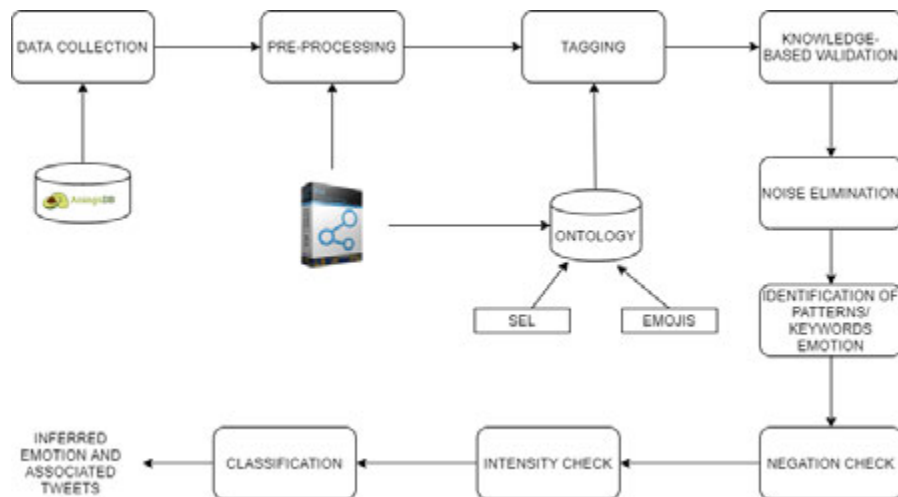


Figure 27. General scheme of the functional architecture of the analysis system.

The description of the functional modules of the analysis process observed in the illustration is as follows:

- **Data collection:** process in charge of consulting from the ArangoDB database the Twitter resources extracted by the SMART-RTVE platform for its later analysis. When these resources are consulted, they are filtered based on the team and journey to be studied and it is obtained the text of the filtered resources.
- **Pre-processing:** module that performs the transformation of the input text into an appropriate format for later indexing. The processes included in the transformation are tokenization, normalization and syntactic labelling of the text. This unit uses the functions offered by the Knowledge Manager to execute these text processing techniques.
- **Tagging:** process that indexes the processed texts according to the defined knowledge (ontology and patterns) through the tool Knowledge Manager for the detection of emotions in them.
- **Knowledge-based validation:** module that checks for valid patterns in the result of the indexing.
- **Noise elimination:** process that matches the patterns that indicate that the text refers to information not relevant to the analysis (information about other events or other categories of team) for later elimination from the process.

- **Identification of patterns/keywords emotion:** part of the system in charge of identifying those patterns found in the text that are associated with a particular emotion and accounting for their occurrence.
- **Negation check:** process of identifying those patterns that detect the presence of an emotion in a denied way, to avoid counting it.
- **Intensity check:** module in charge of identifying the level of strength of the detected emotion, modifying the occurrence value depending on whether the emotional words are accompanied by intensifiers or mitigators.
- **Classification:** process that performs the inference of the emotion corresponding to each tweet, by counting the number of occurrences of each emotion and their weight, being the emotion with the highest associated weight the inferred emotion for that resource. When the emotion has been extracted from each tweet, the percentage of tweets associated with each emotion is determined and the emotion with the highest percentage is established as the global emotion. In the case of the analysis strategy that considers the result (objective context), this percentage is modified according to the result, enhancing positive emotions if the result is favourable or negative emotions if the result is unfavourable.

12.7. Experiments and results

In order to evaluate the results obtained from the emotional classification of the resources analysed, a random sample of 250 tweets that have been collected by the SMART-RTVE platform and analysed by the system has been prepared and provided to three annotators for manual classification within the following categories: Neutral, Joy, Sadness, Anger, Surprise, Dislike and Fear. Based on those tweets that have been labelled with the same category by the 3 scorers, in other words, those that have coincided in the emotion detected, it will be evaluated the classification of the system by calculating the metrics accuracy, precision, recall and F- measure.

Before analysing the value of the metrics obtained for each category, it has to be emphasized the low percentage of coincidence between the evaluators (35%) which indicates the great complexity of this type of classification problem due to its high subjectivity.

The result of the calculation of the accuracy is 65% and the F-measure is 0.54. Firstly, it should be noted that the evaluators have not used the 'Fear' category to label the tweets of the sample, which indicates that this emotion may not adjust to the field in which the analysis is performed. Therefore, a decision could be made to remove this category from the classification problem for system improvement. The F-measure obtained from the same sample if this category is removed is 63%, which means a performance improvement of 9%.

Secondly, another problem encountered is the low precision value in 'Dislike' and 'Sadness', despite the high value in recall, indicating that the quality of the classification for these classes is low. After the analysis of this aspect, it has been concluded that the system presents certain difficulties to distinguish these emotions with regard to the set of negative emotions studied ('Sadness', 'Anger' and 'Dislike').

12.8. Conclusions

After the design and development of the system as well as its later evaluation, it is presented below the conclusions obtained based on the fulfilment of the objectives established for the system.

Firstly, it has been studied and evaluated the different approaches and resources available for the resolution of the engineering problem to be solved in the project, the automatic emotion detection in text, and certain aspects have been detected that have caused several limitations. On the one hand, a lack of development of this field with respect to the existing one for the analysis of sentiment or polarity and, on the other hand, the shortage of available resources for analysis in Spanish (datasets, lexicons, etc.). Despite these problems, an effective solution to the problem has been proposed, which consists of a knowledge-based system based on an ontology composed of the SEL lexicon and a set of emojis, which allows the detection of the categories of emotion defined by Ekman in texts from social networks.

After the definition of the solution, the analysis, design and implementation of the same has been performed. In this, an automatic emotion detection system has been developed for sports communities in which the dimensions of interest for the analysis have been identified (sports event, league, group, team, date of the event, result and emotion). It has been defined a

process for the analysis of the great amount of resources from Twitter where pattern matching techniques and algorithms for the filtering of relevant resources and emotion detection have been applied (both at tweet and global level). In addition, the architecture is based on a Model-View-Controller (MVC) pattern which is suitable for the implementation of the solution through REST type web services and its integration with the existing platform SMART-RTVE. On the other hand, a user interface has been implemented which has been validated by RTVE staff and which allows the visualisation of the analysis results as well as user interaction for the selection of parameters.

Finally, in addition to the verification of the correct functioning of the system by performing requirement validation tests, based on the values obtained from the accuracy and F1 measure, it has been determined that the system has an average performance of over 50%, an accuracy value of 65% and an F-measure of 54%. It should be noted that the value obtained from F1 in the case of eliminating the 'Fear' category due to lack of relevance in the applied domain would be 63% (improving performance by 9%). The system shows a higher performance in the case of positive emotions, 'Surprise' with 85% and 'Joy' with 78%, and in the case of the 'Neutral' category with 71%. However, this performance value can be improved and, therefore, new lines of improvement remain open

Bibliografía

- [1] P. RTVE, « La Fundación Universidad Carlos III y RTVE constituyen la cátedra RTVE-UC3M,» RTVE.ES, 7 Octubre 2015. [En línea]. Disponible en: <https://www.rtve.es/rtve/20151007/fundacion-universidad-carlos-iii-rtve-constituyen-catedra-rtve-uc3m/1234240.shtml>. [Último acceso: 2020 Febrero 4].
- [2] J. Pedraza, «De qué hablamos cuando hablamos de fútbol (en Twitter),» EL PAIS, 6 Mayo 2018. [En línea]. Disponible en: https://elpais.com/deportes/2018/04/30/actualidad/1525080876_301424.html. [Último acceso: 20 Febrero 2020].
- [3] A. Seyeditabari, N. Tabari y W. Zadrozny, «Emotion Detection in Text: a Review,» 2018.
- [4] G. Qiu, X. He, F. Zhang, Y. Shi, J. Bu y C. Chen, «DASA: Dissatisfaction-oriented Advertising based on Sentiment Analysis,» *Expert Systems with Applications*, vol. 37, n° 9, p. 6182–6191, 2010.
- [5] Y. Seki, K. Eguchi, N. Kando y M. Aono, «Multi-document summarization with subjectivity analysis at duc 2005,» de *Document Understanding Conference (DUC)*, 2005.
- [6] E. Riloff, J. Wiebe y W. Phillips, «Exploiting subjectivity classification to improve information extractionl,» de *20th national conference on Artificial intelligence* , 2005.
- [7] . N. Gupta, M. Gilbert y G. Di Fabrizio, «Emotion detection in email customer care,» *Computational Intelligence*, vol. 29, n° 9, p. 489–505, 2013.
- [8] M. Park, C. Cha y M. Cha, «Depressive moods of users portrayed in twitter,» de *ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, 2012.

- [9] J. Bollen, H. Mao y A. Pepe, «Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena,» de *International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, 2011.
- [10] P. Ekman, «An argument for basic emotions,» vol. 6, nº 3-4, p. 169–200, 1992.
- [11] R. Plutchik, «Emotions: A general psychoevolutionary theory,» 1984, p. 197–219.
- [12] J. A. Russell, «A circumplex model of affect,» *Journal of personality and social psychology*, vol. 39, nº 6, pp. 1161-1178., 1980.
- [13] S. Hardik, D. Gosai y H. Gohil, «A review on a emotion detection and recognition from text using natural language processing.,» 2018.
- [14] X. Carreras, I. Chao, L. Padró y M. Padró, «FreeLing: An Open-Source Suite of Language Analyzers,» de *4th International Conference on Language Resources and Evaluation (LREC'04)*, 2004.
- [15] N. 3. documentation, «Natural Language Toolkit,» [En línea]. Disponible en: <http://www.nltk.org/index.html>. [Último acceso: 2020 Marzo 4].
- [16] G. A. Miller, «WordNet: A Lexical Database for English,» *Communications of the ACM*, vol. 38, nº 11, pp. 39-41, 1995.
- [17] The Reuse Company, «KM - Knowledge Manager,» [En línea]. Disponible en: <https://www.reusecompany.com/km-knowledge-manager>. [Último acceso: 4 Marzo 2020].
- [18] S. Greene y P. Resnik, «More than words: Syntactic packaging and implicit sentiment,» de *The 2009 annual conference of the north American chapter of the association for computational linguistics*, 2009.

- [19] E. Cambria, A. Hussain, C. Havasi y C. Eckl, «Affectivespace: Blending common sense and affective knowledge to perform emotive reasoning,» de *WOMSA at CAEPIA*, Sevilla, 2009.
- [20] Swiss center for affective sciences (SCA), *International Survey On Emotion Antecedents And Reactions (ISEAR)*.
- [21] A. Balahur, J. M. Hermida, A. Montoyo y R. Muñoz, «EmotiNet: A Knowledge Base for Emotion Detection in Text Built on the Appraisal Theories,» de *Natural Language Processing and Information Systems - 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011*, Alicante, 2011.
- [22] I. D. Wood y S. Ruder, «Emoji as Emotion Tags for Tweets,» de *Emotion and Sentiment Analysis Workshop*, 2016.
- [23] C. Strapparava y A. Valitutti, «Wordnet affect: an affective extension of wordnet,» de *4th International Conference on Language Resources and Evaluation*, Lisbon, 2002.
- [24] J. W. Pennebaker, M. E. Francis y R. J. Booth, «Linguistic inquiry and word count: Liwc 2007,» Austin, 2007.
- [25] J. Carrillo de Albornoz, L. Plaza y P. Gervás, «SentiSense: An easily scalable concept-based affective lexicon,» *LREC*, vol. 12, pp. 3562-3567, 2012.
- [26] G. Sidorov, S. Miranda-Jiménez, F. ViverosJiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. SuárezGuerra, A. Treviño y J. Gordon, «Empirical study of machine learning based,» de *Mexican international conference on Artificial intelligence*, 2012.
- [27] W. Y. Zou, R. Socher, D. Cer y C. D. Manning, «Bilingual word embeddings for phrase-based machine translation,» de *2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

- [28] J. Turian, L. Ratinov y Y. Bengio, «Word representations: a simple and general method for semi-supervised learning,» de *48th annual meeting of the association for computational linguistics*, 2010.
- [29] K. Sailunaz, M. Dhaliwal, J. Rokne y R. Alhajj, «Emotion Detection from Text and Speech - A Survey,» de *Social Network Analysis and Mining*, vol. 8, Springer, 2018, pp. 1-26.
- [30] S. Badugu y M. Suhasini, «Emotion Detection on Twitter Data using Knowledge Base Approach,» de *Internal Journal of Computer Applications*, 2017.
- [31] A. Esuli y F. Sebastiani, «SENTIWORDNET: A Publicly Available Lexical Resource,» *LREM*, vol. 6, pp. 417-422, 2006.
- [32] F. M. Shah, A. S. Reyadh, A. I. Shaafi, S. Ahmed y F. T. Sithil, «Emotion Detection from Tweets using AIT-2018 Dataset,» de *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, Dhaka, Bangladesh, 2019.
- [33] S. M. Mohammad, F. Bravo-Marquez, M. Salameh y S. Kiritchenko, «SemEval-2018 Task 1: Affect in Tweets,» de *12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, Louisiana,, 2018.
- [34] S. Poria, A. Gelbukh, . A. Hussain, D. Das y S. Bandyopadhyay, «Enhanced SenticNet with Affective Labels for Concept-based Opinion Mining,» de *IEEE Intelligent Systems*, vol. 28, 2013, p. 31–38.
- [35] G. Blázquez Gil, A. Berlanga de Jesús y J. M. Molina López, «Combining Machine Learning Techniques and Natural Language Processing to infer Emotions Using Spanish Twitter Corpus,» *Communications in Computer and Information Science*, n° 365, pp. 149-157, 2013.

- [36] C.-H. Wu, Z.-J. Chuang y Y.-C. Lin, «Emotion recognition from text using semantic labels and separable mixture models,» *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 5, pp. 165-183, 2006.
- [37] IBM Cloud, «Tone Analyzer,» [En línea]. Disponible en: <https://cloud.ibm.com/apidocs/tone-analyzer>. [Último acceso: 10 Marzo 2020].
- [38] ParallelDots, «Emotion Analysis,» [En línea]. Disponible en: <https://www.paralleldots.com/emotion-analysis>. [Último acceso: 10 Marzo 2020].
- [39] SDKS.IO, «MoodPatrol - Emotion Detection from Text,» [En línea]. Disponible en: <https://www.sdk.sio/SDK/View/moodpatrol-emotion-detection-from-text-2>. [Último acceso: 10 Marzo 2020].
- [40] Github. [En línea]. Disponible en: <https://github.com/about>. [Último acceso: 2020 Marzo 17].
- [41] Microsoft, «Visual Studio 2019,» [En línea]. Disponible en: <https://visualstudio.microsoft.com/es/vs/>. [Último acceso: 17 Marzo 2020].
- [42] Microsoft, «ASP.NET overview,» [En línea]. Disponible en: <https://docs.microsoft.com/en-us/aspnet/overview>. [Último acceso: 17 Marzo 2020].
- [43] Microsoft, «Paseo por el lenguaje C#,» [En línea]. Disponible en: <https://docs.microsoft.com/es-es/dotnet/csharp/tour-of-csharp/>. [Último acceso: 17 Marzo 2020].
- [44] Angular. [En línea]. Disponible en: <https://angular.io/>. [Último acceso: 17 Marzo 2020].
- [45] Highcharts, [En línea]. Disponible en: <https://www.highcharts.com/>. [Último acceso: 17 Marzo 2020].

- [46] ArangoDB, [En línea]. Disponible en: <https://www.arangodb.com/>. [Último acceso: 17 Marzo 2020].
- [47] MySQL. [En línea]. Disponible en: <https://www.mysql.com/>. [Último acceso: 17 Marzo 2020].
- [48] P. Kruchten, «The 4+1 View Model of Architecture,» *IEEE Software*, vol. 12, nº 6, pp. 42-50, 1995.
- [49] Universidad de Alicante, «Modelo vista controlador (MVC),» [En línea]. Disponible en: <https://si.ua.es/es/documentacion/asp-net-mvc-3/1-dia/modelo-vista-controlador-mvc.html>. [Último acceso: 1 Abril 2020].
- [50] Newtonsoft, «Json.NET,» [En línea]. Disponible en: <https://www.newtonsoft.com/json>. [Último acceso: 31 Mayo 2020].
- [51] «Angular Material,» [En línea]. Disponible en: <https://material.angular.io/>. [Último acceso: 31 Mayo 2020].
- [52] J. M. Heras, «Precision, Recall, F1, Accuracy en clasificación,» [En línea]. Disponible en: <https://iartificial.net/precision-recall-f1-accuracy-en-clasificacion/>. [Último acceso: 31 Mayo 2020].
- [53] «Grigori SIDOROV,» [En línea]. Disponible en: <http://www.cic.ipn.mx/~sidorov/#SEL>. [Último acceso: 20 Enero 2020].
- [54] Diario Oficial de la Unión Europea, «REGLAMENTO (UE) 2016/679,» 27 Abril 2016. [En línea]. Disponible en: <https://www.boe.es/doue/2016/119/L00001-00088.pdf>. [Último acceso: 31 Mayo 2020].
- [55] BOE, «Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.,» 6 Diciembre 2018. [En línea]. Disponible en: <https://www.boe.es/eli/es/lo/2018/12/05/3/con>. [Último acceso: 18 Mayo 2020].

- [56] Twitter, «Política de Privacidad,» 1 Enero 2020. [En línea]. Disponible en: <https://cdn.cms-twdigitalassets.com/content/dam/legal-twitter/site-assets/privacy-page-gdpr/pdfs/Twitter-Privacy-Policy-ES.pdf>. [Último acceso: 31 Mayo 2020].
- [57] BOE, «Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia.,» 22 Abril 1996. [En línea]. Disponible en: <https://www.boe.es/buscar/act.php?id=BOE-A-1996-8930>. [Último acceso: 28 Mayo 2020].
- [58] B. Adsuara, «¿Puede ser un tuit objeto de propiedad intelectual?,» 23 Noviembre 2015. [En línea]. Disponible en: https://blogs.elconfidencial.com/tecnologia/menos-tecnologia-y-mas-pedagogia/2015-11-23/puede-ser-un-tuit-objeto-de-propiedad-intelectual_1102076/. [Último acceso: 28 Mayo 2020].
- [59] A. Garcia Zarzalejos, «La transformación digital de los medios de comunicación,» El Confidencial, 5 Agosto 2018. [En línea]. Disponible en: https://www.elconfidencial.com/comunicacion/2018-08-05/medios-liquidos-transformacion-sector_1600924/. [Último acceso: 18 Mayo 2020].
- [60] D. Martínez, «Transformación digital en los medios de comunicación,» Observatorio eCommerce, [En línea]. Disponible en: <https://observatorioecommerce.com/transformacion-digital-medios-comunicacion/>. [Último acceso: 18 Mayo 2020].
- [61] Future Today Institute, «2018 Tech Trends For Journalism and Media,» 2018.
- [62] indeed, «Salarios para empleos de Jefe de proyecto en España,» 13 Mayo 2020. [En línea]. Disponible en: <https://www.indeed.es/salaries/jefe-de-proyecto-Salaries>. [Último acceso: 23 Mayo 2020].

- [63] indeed, «Salarios para empleos de Programador/a ASP.net en España,» [En línea]. Disponible en: <https://www.indeed.es/salaries/programador-asp.net-Salaries>. [Último acceso: 23 Mayo 2020].
- [64] A. Moreno Ortiz, Á. Pérez Pozo y S. Torres Sánchez, «Sentitext: sistema de análisis de sentimientos para el español,» *Procesamiento del Lenguaje Natural*, nº 45, pp. 297-298, Septiembre 2010.
- [65] bogdani, «Getting Started with Sentiment Analysis,» 9 Noviembre 2016. [En línea]. Disponible en: <https://nlpforhackers.io/sentiment-analysis-intro/>.
- [66] J. Pedraza, «De qué hablamos cuando hablamos de fútbol (en Twitter),» *EL PAIS*, 6 Mayo 2018.
- [67] F. M. Shah, A. S. Reyadh, A. I. Shaaf, S. Ahmed y F. T. Sithil, «Emotion Detection from Tweets using AIT-2018 Dataset,» *5th International Conference on Advances in Electrical Engineering (ICAEE)*, 2019.
- [68] M. Rouse, «knowledge-based systems (KBS),» 31 Mayo 2018. [En línea]. Disponible en: <https://searchcio.techtarget.com/definition/knowledge-based-systems-KBS>. [Último acceso: 24 Marzo 2020].
- [69] S. Badugu y M. Suhasini, «Emotion Detection on Twitter Data using Knowledge Base Approach,» *Internal Journal of Computer Applications*, vol. 162, pp. 28-33, 2017.
- [70] G. Gutiérrez, L. Margain, C. de Luna, A. Padilla, J. Ponce, J. Canul y A. Ochoa, «A Sentiment Analysis Model: To Process Subjective Social Corpus through the Adaptation of an Affective Semantic Lexicon,» de *Human-Inspired Computing and Its Applications*, 2014, pp. 233-244.
- [71] Meaning cloud, «Análisis de Sentimiento,» [En línea]. Disponible en: <https://www.meaningcloud.com/es/productos/analisis-de-sentimiento>. [Último acceso: 10 Marzo 2020].

[72] indeed, [En línea]. Disponible en: <https://www.indeed.es/>. [Último acceso: 21 Mayo 2020].

[73] Twitter, «Developer Terms. More about restricted uses of the Twitter APIs,» [En línea]. Disponible en: <https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases>. [Último acceso: 27 Mayo 2020].